PRDM9 – New Locus Identified by GWAS for Skin Intrinsic Fluorescence

Zeya Chen

Supervised by Prof. Andrew Paterson

2021-04-28

Abstract

Intro

Skin intrinsic fluorescence (SIF), a non-invasive skin glycation measure, has previously shown to be influenced by genetics. With contemporary imputation technology, we suspect new genetic variations can be identified, thus we performed genome wide association studies (GWAS) of SIF.

Method

1082 type 1 diabetes (T1D) patients were selected from the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications study.

Result

rs62342757(C>T, Minor Allele Frequency = 0.38, Imputation Quality = 0.94) from chromosome 5 is associated with SIF ($\beta \pm SE = -0.040 \pm 0.007$, $p = 3.9 * 10^{-8}$, Semi-Partial $R^2 = 0.018$), belongs to PR domain zinc finger protein 9 (*PRDM9*). We then ran a regional conditional GWAS on chromosome 5 & 8 to check for secondary signals, but no secondary signals were identified.

Conclusion

We Identified an association between SNPs near *PRDM9* and SIF in T1D patients. Specific variants in *PRDM9* are known to affect the rate of meiotic recombination. Replication is required to confirm this association.

Keywords: Skin Intrinsic Fluorescence; Genome-wide association study; Diabetes; *PRDM9*.

Introduction

Diabetes Background

Approximately 10% of Canadian adults have diabetes (**Public Health Canada, 2017, 1**). Diabetes is caused by malfunction in insulin production. Since insulin helps glucose enter into cells, insufficiency or absence of insulin will cause glucose to build up in blood. Long term of high blood glucose level will result in glycation (glucose binding with other proteins) and lead to microvascular and macrovascular complications like retinopathy, nephropathy, ischemic heart disease, and peripheral vascular disease (**Cade, 2008, 2**).

Variable of Interest

Skin intrinsic fluorescence is a non-invasive measure for advanced glycated end products (AGE) in the skin, it is performed by emitting ultraviolet/blue light and measuring emission from the skin. Participants put their arm on Scout DS, and the machine takes measurements from their forearm as shown in Figure 1.

Figure 1: SIF Measurement
(https://www.diabetesnet.com/diabetes-technology
/insulin-pumps/future-pumps/scout-ds-by-veralight/)

Predictor of Interest and Purpose

In genome wide association studies (GWAS), the aim is to discover genetic variations that influence an outcome. However, genes cannot be quantified, while

single nucleotide polymorphisms (SNPs) can. SNPs are base pairs on the genome that vary between individuals. SNPs can occur within a genes' regulating region or anywhere else. It is not currently cost effective to sequence the entire genome for all subjects in the study and SNPs are highly correlated, hence small portions of SNPs are genotyped while remaining SNPs are imputed.

Rationale and Hypothesis

N-acetyltransferase 2 (NAT2) gene was initially identified to influence SIF (**Eny, 2014, 3**), using the HapMap II imputation reference (sample size 60, sites 2,542,916). rs7533564 from chromosome 1 was later identified (**Roshandel, 2016, 4**), with imputation reference from the 1000 Genomes Phase 1 Ver. 3 (sample size 1092, sites 28,975,367). TOPMed Ver. r2 (sample size 97,256, sites 308,107,085) imputation reference surpasses predecessors not only in quantity but also in quality, since all samples were sequenced for their whole genome using high coverage approaches called next generation sequencing (NGS). NGS is more accurate so it improves imputation quality of rare variations (**Taliun, 2019, 5**).

Therefore, with state-of-the-art imputation reference TOPMed Ver. r2, we are expecting to discover additional genetic variations that influence SIF, and the hypothesis is: With TOPMed imputation reference, we will find at least one SNP that is associated with skin intrinsic fluorescence in type 1 diabetes patients with European ancestry.

Method

Data

Participants

1082 participants with type 1 diabetes (T1D) of European descent were selected from Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications (DCCT/EDIC) study. DCCT started in 1983 and finished recruiting in 1989, the cohort size is 1441 (**DCCT Group, 1993, 6**). Only T1D patients who passed A lot of examinations/tests were chosen. DCCT transitioned to EDIC in 1994, and 1185 participants had SIF measured in 2009 – 2010 (**Cleary, 2013, 7**). After removing Non-Europeans or missing GWAS data participants, we end up with 1082 participants.

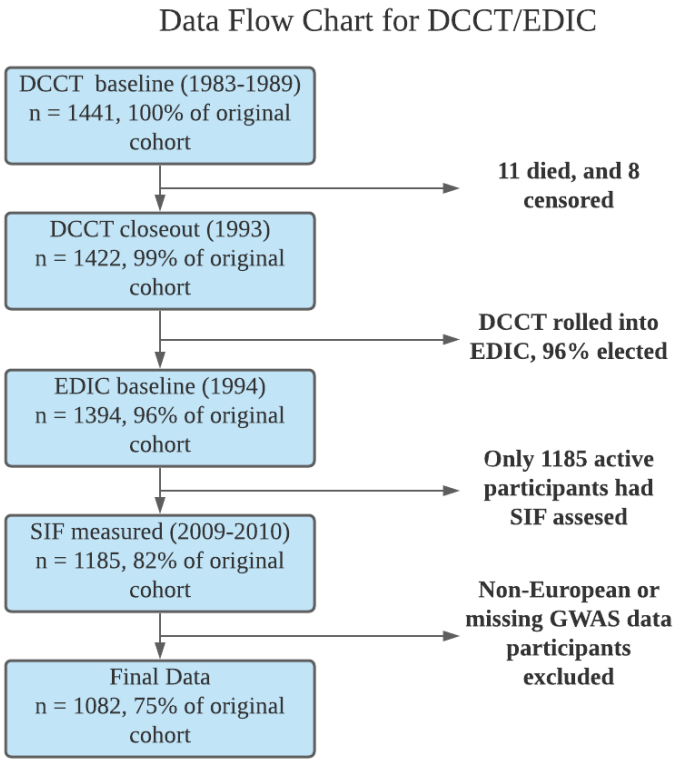


Figure 2: Data Flow Chart

SIF Measure

SIF was measured 16 years after the closeout of DCCT (**Cleary, 2013, 7**). A total of 5 LED emission levels were measured, and 15 measurements were recorded. The main outcome SIF1 LED level is at 375nm, emission range is between 435nm and 655nm, dimensionless excitation factor is 0.6 and emission exponents factor is 0.2. The rest of measurements LED level are at 405nm, 416nm, 435nm, 456nm. The main outcome SIF1 is all positive and right-skewed with arbitrary unit, so a natural log-transform is necessary. Figure 3 below showed SIF1 achieved normality after natural log transformation.

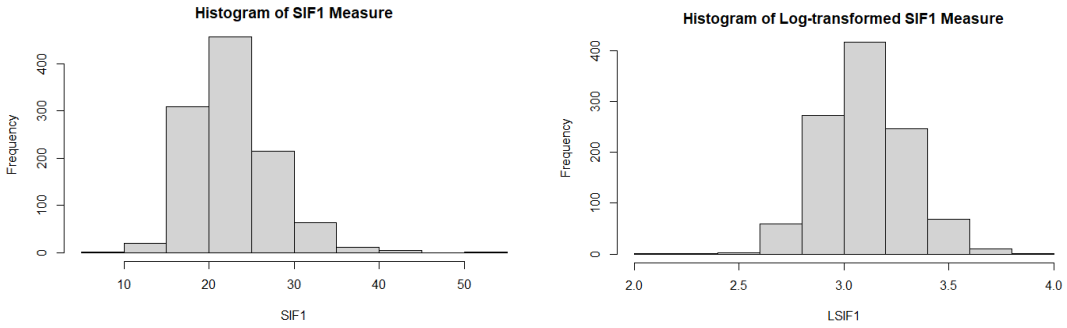


Figure 3: Histograms of SIF1 and Log-transformed SIF1

SNPs Imputation

841,342 SNPs are genotyped using an Illumina Infinium 1M beadchip assay (https://www.illumina.com/science/technology/microarray.html) after quality control. 10,235,659 SNPs with minor allele frequency > 0.5% and information criteria > 0.3 was imputed using the TOPMed Ver.r2 reference (**Taliun, 2019, 5**) (**Das, 2016, 8**) (**Fuchsberger, 2014, 9**). Imputation provides the expected number of minor alleles at each SNP by predicting genotype probability. In order to

preserve the full power of imputation, the expected number of minor alleles is kept in dosage form.

Covariates

Other covariates are included to better distinguish variations in SIF explained by genetics. The covariates included are age, sex, smoking years, skin color, geographic location indicator as proxy for sun-exposure, mean kidney function measure(eGFR), and time weighted HbA1c measure. Smoking years was divided into three variables. Smoking Status 1 to 5 variable is calculated by summing smoking status (smoker = 1, non-smoker = 0) in 1-5 years of EDIC. Smoking Status 6 to 10 variable is calculated by summing smoking status (smoker = 1, non-smoker = 0) in 6-10 years of EDIC. Smoking Status 11 to 16 variable is calculated by summing smoking status (smoker = 1, non-smoker = 0) in 11-16 years of EDIC.

Genome Wide Association Studies

As technology advances, computing power increases and genome sequencing cost drops. GWAS, introduced in 2007, has gradually became the new method to study genetic association. Compared to old era method like linkage analysis, GWAS is time-saving and cost-effective. In statisticians' point of view, GWAS is a looped regression model, it loops through all the SNPs in the study. Phenotype is the dependent variable and can be categorical or continuous. While SNP, other covariates, and possibly interactions are used as independent variables.

Defining the linear regression at each SNP

Variables in the study:

- $\boldsymbol{Y}$: SIF measure - continuous
- $\boldsymbol{G}$: Expected Number of Minor Alleles at Selected SNP – continuous
- $\boldsymbol{X_2}$: Age – continuous
- $\boldsymbol{X_3}$: Sex – categorical
- $\boldsymbol{X_4}$: Smoking Status (1 to 5 years) – continuous
- $\boldsymbol{X_5}$: Smoking Status (5 to 10 years) – continuous
- $\boldsymbol{X_6}$: Smoking Status (11 to 16 years) – continuous
- $\boldsymbol{X_7}$: Skin Color – continuous
- $\boldsymbol{X_8}$: Geographic Location Indicator – categorical
- $\boldsymbol{X_9}$: Average Kidney Function – continuous
- $\boldsymbol{X_{10}}$: Time Weighted HbA1c – continuous

Let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_{1082} \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & \cdots & X_{1,10} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & X_{1082,10} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{10} \end{bmatrix}, \qquad e = \begin{bmatrix} e_1 \\ \vdots \\ e_{1082} \end{bmatrix}$$

Linear regression model is specified as:

$$Y = \beta_0 + \beta_1 G + \sum \beta_i X_i + e \; = \; X\beta \; + \; e$$

β_0 is the intercept, β_1 is the genotype effect of selected SNP, $\beta_{i,i>1}$ are effects of covariates, and e is the residual term. Additive coding is used for genotype effect because it suits imputed data and gives larger power.

Estimate and variance is computed using the following maximum likelihood estimator:

$$\hat{\beta} = (X'X)^{-1}X'Y,\ \ \widehat{\sigma_{\beta}} = (X'X)^{-1}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)n^{-1}$$

P-value of each β_i is from t-test with hypothesis:

$$H_0: \beta_i = 0, \qquad vs. \qquad H_a: \beta_i \neq 0$$

Test statistics is given by:

$$t = \hat{\beta}_i se\left(\hat{\beta}_i\right)^{-1} \sim t_{1070}$$

$\hat{\beta}_i$ is i[th] element of $\hat{\beta}$, and $se\left(\hat{\beta}_i\right)$ is i[th] diagonal element of $\widehat{\sigma_{\beta}}$. The degree of freedom is calculated by $df = n - k - 1 = 1082 - 11 - 1 = 1070$.

We loop the above regression over ~ 10 million SNPs, and record genotype effect (β_1) estimate, standard error, p-value.

Checking GWAS Validity

Under the hypothesis that no genetic variation influences SIF, p-values are uniformly distributed. Quantile-quantile (QQ) plot and histogram can visually check p-values' deviation from uniform distribution. If GWAS detects signals, we are expected to see deviations on the right tail of the QQ-plot and minor bump on the histogram near p = 0. Other patterns indicate GWAS is not reliable since most p-values are not uniformly distributed. Genomic Inflation factor λ is a quantitative checker for type 1 inflation, it is computed by the median of observed chi square statistics over empirical chi square 1 df distribution, and chi square statistics are from squared inverse standard normal transformed p-values.

SNP Selection and Genome Wide Significance

Bonferroni correction is routinely used to determine per-comparison significance level in multiple hypothesis testing, but not applicable to GWAS. Independent test assumption does not hold since nearby SNPs are correlated.

$5 * 10^{-8}$ was proven to be the most effective significance level with a whole genome simulation (**Dudbridge, 2008, 10**). Hence, SNPs with p-value less than $\alpha = 5 * 10^{-8}$ are selected.

Checking Regression Validity

Regression model is the basis for GWAS, so we need to check regression assumptions. SNPs identified are clustered by gene region, checking all SNPs is tedious. Thus, checking regression assumptions of most significant SNPs from each region is sufficient. We will check SNPs correlation with covariates and residual distribution.

Regional Conditional Analysis

Regional conditional analysis detects independent secondary genetic variation, it is similar to GWAS construction. As the name suggests, regional implies that analysis is performed on chromosomes with identified gene region, conditional implies containing primary signal as covariates.

All variables stay the same as defined in GWAS. We define G_* as the most significant SNP from region. The linear regression model became:

$$Y = \beta_0 + \beta_1 G + \sum \beta_i X_i + \beta_* G_* + e$$

Computational method stays the same as stated in GWAS. We loop this regression model over selected chromosomes and the rest of the procedure follows GWAS.

Unsupervised Method on Covariates and SIFs

By the multivariate nature of data, multiple-linear regression will not capture all relations. Unsupervised learning algorithms can possibly bring data mining to a next level.

Principal Component Analysis on SIFs

Principal component analysis (PCA) is used to reduce data dimension so data can be studied on a lower dimensional space. It is computed by eigen decomposing variance-covariance matrix of data. Eigenvectors are the principal components, ranked by their eigenvalue magnitude in descending order. Percentage of variance explained by the eigenvector is computed by its eigenvalue over summation of all eigenvalues. Since SIF has 15 measurements, we can apply PCA on measurements and use scree plot and biplot to analyze SIF data structure.

K-mean clustering of SIF and Covariates

K-mean is an iterative non-parametric data clustering method. It is constructed by a starting vector for cluster means and a while loop with two steps. While loop condition is: the normed difference between current cluster means and previous

iterate cluster means converged by a pre-determined threshold ϵ; or reaching maximum iterations. The two steps are:

1. Send each observation to the cluster that has least Euclidean distance between mean and observation.
2. Compute sample means of all current clusters and update the cluster mean vector.

We can use k-mean to cluster dataset made of covariates and SIF, comparing the cluster with genotype categories.

Partition Around Medoids clustering of SIF and Covariates

Partition Around Medoids (PAM) is also known as k-median. All procedures stay the same, except medoids (multivariate median) are used instead of mean.

Software

Software used in this research include:

1. TOPMed imputation server Version r2 for SNP imputation (https://imputation.biodatacatalyst.nhlbi.nih.gov/#!).
2. SNPTEST 2.5 on Sickkids high performance computing cluster for GWAS (https://ccm.sickkids.ca/high-performance-computing/).
3. R 4.0.4 on personal computer for validity checking, plotting, and clustering

Results

Primary GWAS Results

From Table 1, we can see that the majority of the participants are in their 50s in 2010. There were slightly more males than females, and they are mostly from northern latitude.

Table 1: Descriptive Statistics of Covariates and Trait in DCCT/EDIC.

	Mean	Standard Error
log SIF1	3.1	0.2
Age (years)	51.51	7
Smoking Status 1 to 5 (years)	0.87	1.7
Smoking Status 6 to 10 (years)	0.73	1.6
Smoking Status 11 to 16 (years)	0.75	1.8
Skin Color	261.7	41.3
Mean eGFR (mL/min/1.73 m2)	108.7	10.1
Time Weighted HbA1c (%)	8.2	0.9
Sex	Male:579	Female:503
Geographic Location	North:800	South:282

- Mean eGFR stands for mean estimated glomerular filtration rate, it is a measure for kidney function.

From Figure 4, we can clearly observe two signals that were genome wide significant, one from chromosome 8, the other from chromosome 5. Chromosome 8 signal is the previously identified NAT2 gene (**Eny, 2014, 3**), and chromosome 5 signal is newly identified.

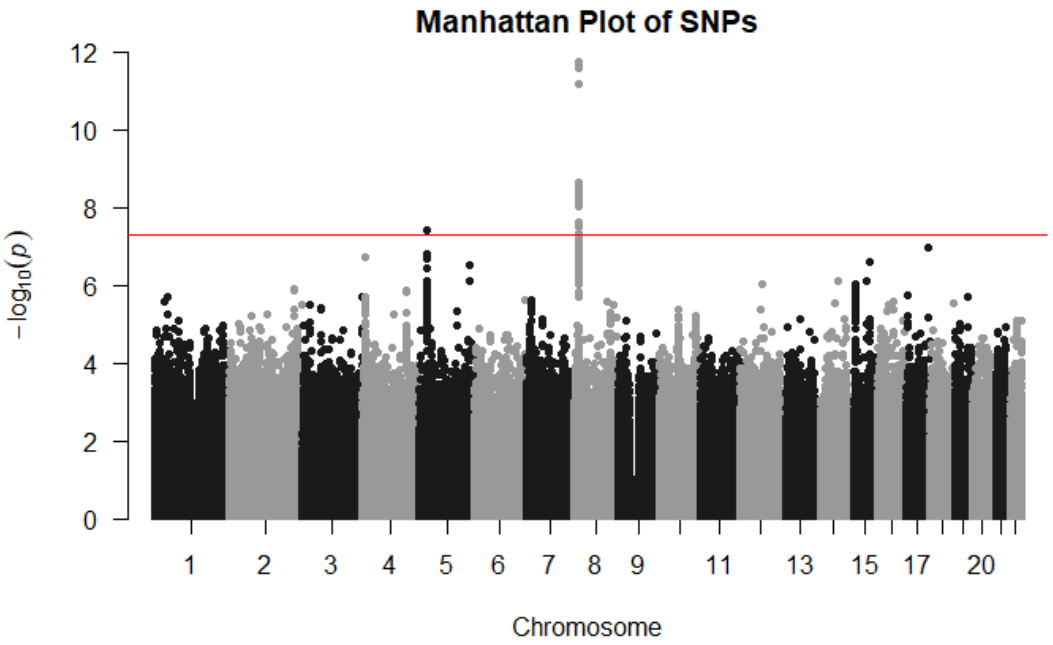


Figure 4: Manhattan Plot of DCCT SIF1 GWAS. Produced using r package "qqman" (**Turner, 2014, 11**). Red line: genome wide significance level $5 * 10^{-8}$. X axis: SNPs physical location on each chromosome. Y axis: Log base 10 transform P-value for each SNPs.

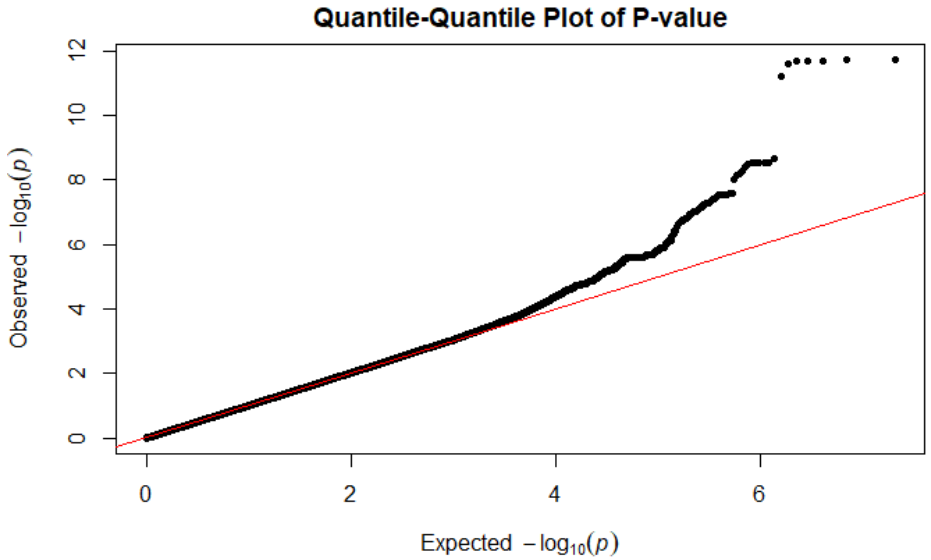


Figure 5: QQ Plot of DCCT SIF1 GWAS p-values. Produced using r package "qqman" (**Turner, 2014, 11**). Red line represents identity function. Y-axis is the same as Manhattan plot, and X-axis is log base 10 transformed rank quantile.

From Figure 5, we can see that the majority of p-values stay on the identity line while some deviation shows up on the right side. The QQ-plot pattern indicates that GWAS did identify signals and the result is reliable.

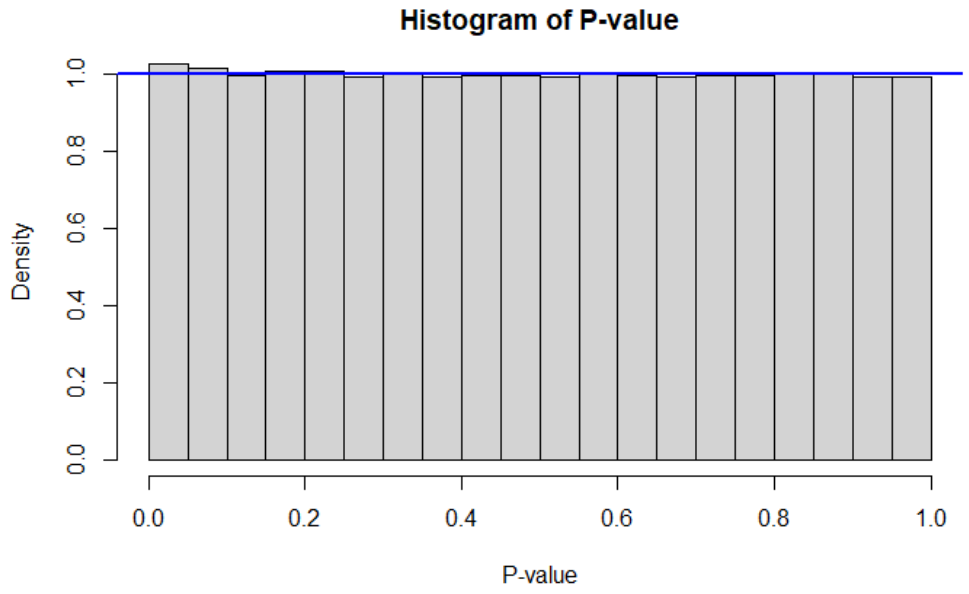


Figure 5: Histogram of DCCT SIF1 GWAS p-values. Blue line represents uniform (0,1) density function.

The histogram shows minor bump near p-value = 0, while the rest of the density are close to uniform. Genome Inflation factor λ is 1.009, so type 1 error inflation is negligible. All three checkers imply that our result is reliable, and selected SNPs are shown below:

Table 2: DCCT SIF1 GWAS Signals.

Chromosome	Position	Imputation Quality	Minor Allele Frequency	P-value	Effect Estimate	Effect Standard Error
8	18415371	1.00	0.22	1.80E-12	0.059	0.0083
8	18415125	0.99	0.23	1.80E-12	0.059	0.0083
8	18415025	0.99	0.23	2.07E-12	0.059	0.0083
8	18414867	0.99	0.23	2.11E-12	0.059	0.0083
8	18414928	0.99	0.23	2.12E-12	0.059	0.0083
8	18415790	0.99	0.23	2.54E-12	0.059	0.0083
8	18414956	0.99	0.23	6.41E-12	0.058	0.0083
8	18383997	1.00	0.15	2.27E-09	-0.058	0.0097
8	18402366	0.99	0.28	2.78E-09	0.046	0.0077
8	18402516	1.00	0.28	2.91E-09	0.046	0.0077
8	18402845	1.00	0.28	2.97E-09	0.046	0.0077
8	18402921	1.00	0.28	2.97E-09	0.046	0.0077
8	18403088	1.00	0.28	2.97E-09	0.046	0.0077
8	18403983	1.00	0.28	3.16E-09	0.046	0.0077
8	18402503	0.99	0.28	4.08E-09	0.046	0.0077

8	18396999	1.00	0.16	5.19E-09	-0.056	0.0094
8	18390099	1.00	0.15	6.67E-09	-0.057	0.0098
8	18401856	1.00	0.28	6.81E-09	0.045	0.0077
8	18392865	1.00	0.15	9.65E-09	-0.057	0.0098
8	18387939	1.00	0.19	2.49E-08	-0.049	0.0087
8	18405213	1.00	0.27	2.72E-08	0.044	0.0078
8	18405332	0.99	0.27	2.77E-08	0.044	0.0078
8	18405974	1.00	0.27	2.87E-08	0.043	0.0078
8	18405602	0.99	0.28	2.92E-08	0.043	0.0078
8	18405351	0.99	0.28	2.93E-08	0.043	0.0078
8	18404610	0.99	0.28	2.93E-08	0.043	0.0078
8	18404921	0.99	0.28	2.94E-08	0.043	0.0078
8	18405008	0.99	0.28	3.20E-08	0.043	0.0078
5	23575628	0.94	0.38	3.89E-08	-0.040	0.0073
5	23574831	0.94	0.38	3.99E-08	-0.040	0.0073
8	18387243	1.00	0.20	4.40E-08	-0.047	0.0086
8	18409473	1.00	0.15	4.87E-08	-0.053	0.0096

Row highlighted in green are signals that were not identified previously.

From Table 2, except the two SNPs highlighted in green, all SNPs from chromosome 8 with base pair number 18380000-18420000 (Build 38) are in previously identified NAT2 region. We will disregard these signals since this locus was previously reported (**Eny, 2014, 3**).

SNPs highlighted in green are 797 base pairs apart and share similar effect and standard error (-0.040, 0.0073). rs62342757(C>T, Minor Allele Frequency = 0.38, Imputation Quality = 0.94) is the SNP with base pair number 23575628. rs62342757 is our new signal since its p-value is $3.9 * 10^{-8} < 5 * 10^{-8}$, and it explains 1.8% of the total variation in LSIF1. T is minor allele for rs62342757, so effect can be interpreted as:

Controlling for other covariates, T1D patients with one more allele T is expected to have their logged SIF1 measure decreased by 0.04 unit.

From Figure 6, rs62342757 is in PR domain zinc finger protein 9 (*PRDM9*) gene regulating region.

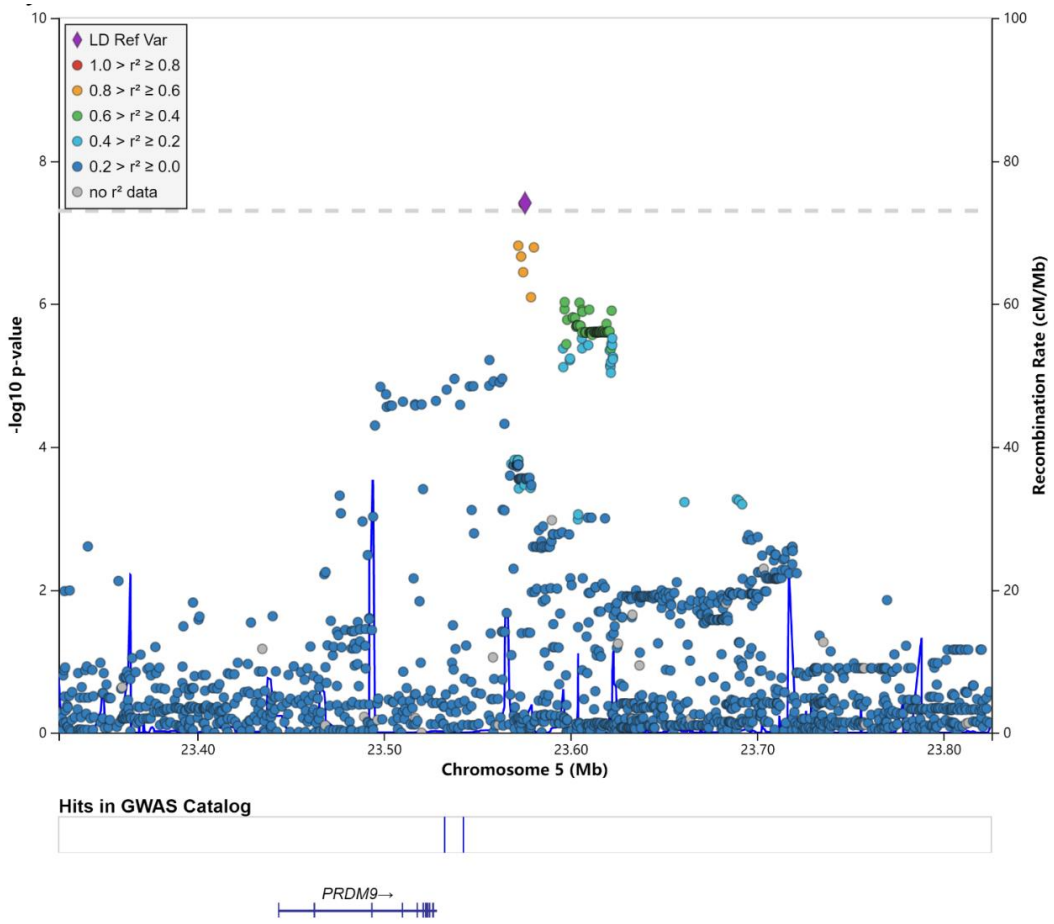


Figure 6: Locus Zoom Plot **(Pruim, 2010, 12)** on rs62342757.

Gray line is genome wide significance level $5 * 10^{-8}$. LD Ref Var: linkage disequilibrium reference variation. Mb: Mega base pairs. *PRDM9*: PR domain zinc finger protein 9 gene. cM: centimorgan. Colors represent correlation between colored allele and reference allele.

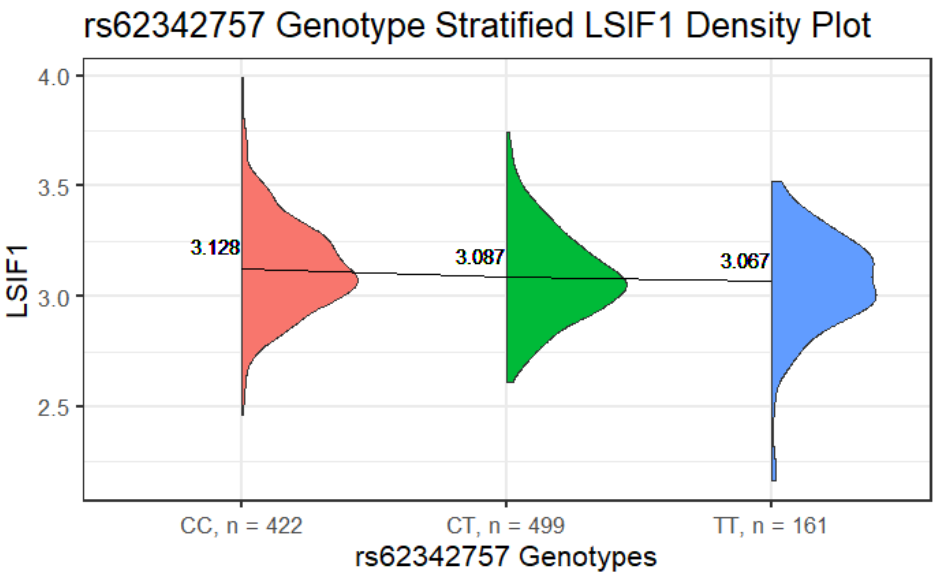


Figure 7: rs62342757 Genotype Stratified Density Plot

LSIF1: Logged-SIF1. Black lines are linear interpolations between two genotypes mean Logged SIF1 measure. Black numbers are genotypes mean Logged-SIF1 measure. n is sample size for each genotype.

From Figure 7, we can observe 0.041 unit of mean level Logged SIF1 decrease when T allele counts increased from 0 to 1 (genotype CC to CT), and 0.02 unit of mean level Logged SIF1 decrease when T allele increased from 1 to 2 (genotype CT to TT). This does not strictly follow the regression result, and it is possibly caused by enforcing continuous dosage form data into categorical data and losing power, and lack of inclusion of covariates.

Table 3: Association of rs62342757 with SIFs in DCCT/EDIC

LED level (Emission range)	SIF (kx, km)	Mean ± SD	β ± SE	p-value
375nm (435-655nm)	SIF1 (Kx0.6 km0.2)	3.10 ± 0.20	-0.040 ± 0.007	3.90E-08
	SIF2 (Kx0.8 Km0.2)	3.20 ± 0.25	-0.047 ± 0.008	3.35E-08
	SIF3 (Kx0.4 Km0.7)	2.62 ± 0.19	-0.035 ± 0.007	1.57E-06
405nm (440-655nm)	SIF4 (Kx0.6 Km0.2)	2.11 ± 0.23	-0.043± 0.009	3.99E-07
	SIF5 (Kx0.8 Km0.2)	2.11 ± 0.23	-0.047 ± 0.008	1.28E-07
	SIF6 (Kx0.9 Km0.0)	2.27 ± 0.24	-0.047 ± 0.008	9.80E-08
416nm (451-655nm)	SIF7 (Kx0.8 Km0.2)	1.84 ± 0.24	-0.047 ± 0.009	3.48E-07
	SIF8 (Kx0.9 Km0.0)	2.00 ± 0.24	-0.048 ± 0.009	2.68E-07
	SIF9 (Kx0.4 Km0.9)	1.28 ± 0.23	-0.043 ± 0.009	2.69E-06
	SIF10 (Kx0.9 Km0.0)	1.59 ± 0.25	-0.049 ± 0.010	7.63E-07

435nm (470-655nm)	SIF11 (Kx0.4 Km0.8)	1.02 ± 0.25	-0.046 ± 0.010	3.61E-06
	SIF12 (Kx0.4 Km0.9)	0.94 ± 0.24	-0.046 ± 0.010	3.73E-06
456nm (491-655nm)	SIF13 (Kx0.9 Km0.0)	0.68 ± 0.24	-0.047 ± 0.010	2.84E-07
	SIF14 (Kx0.4 Km0.8)	0.36 ± 0.23	-0.043 ± 0.009	2.06E-06
	SIF15 (Kx0.4 Km0.9)	0.28 ± 0.23	-0.042 ± 0.009	2.35E-06

Kx: dimensionless excitation factor. Km: emission exponents factor at 0.2.

From Table 3, only SIF1 and SIF2 showed a genome wide significant p-value in 15 SIF measures.

Secondary Results

Checking rs62342757 Regression Model Assumption

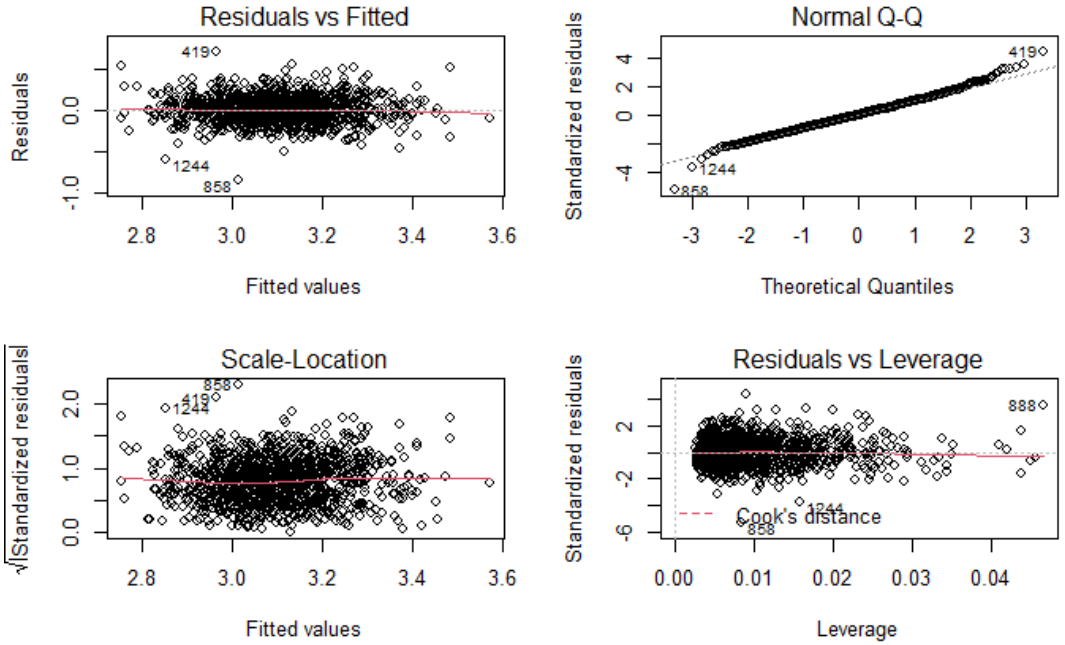
Since covariates were selected to be uncorrelated and some covariates cannot be genetically influenced, we only need to check whether rs62342757 is associated with skin tone, mean kidney function, and time weighted HbA1c.

Table 4: Regression of rs62342757 with covariates in DCCT/EDIC

	β ± SE	P-value
Skin Color	2.29 ± 1.86	0.22
Mean eGFR (mL/min/1.73 m2)	0.31 ± 0.46	0.49
Time Weighted HbA1c (%)	0.006 ± 0.04	0.89

rs62342757 is not associated with any covariates, so there is no multicollinearity concern.

Figure 8: Residual Plots of SIF1 Regression Model with rs62342757



From Figure 8, there is no trend and residuals are evenly spread out, and no significant leverage point. Finally, normal QQ plot shows deviations on two sides but majority lay on the identity line, implying residuals are normally distributed. Hence, all regression assumptions are checked, results from regression are reliable.

Regional Conditional Results

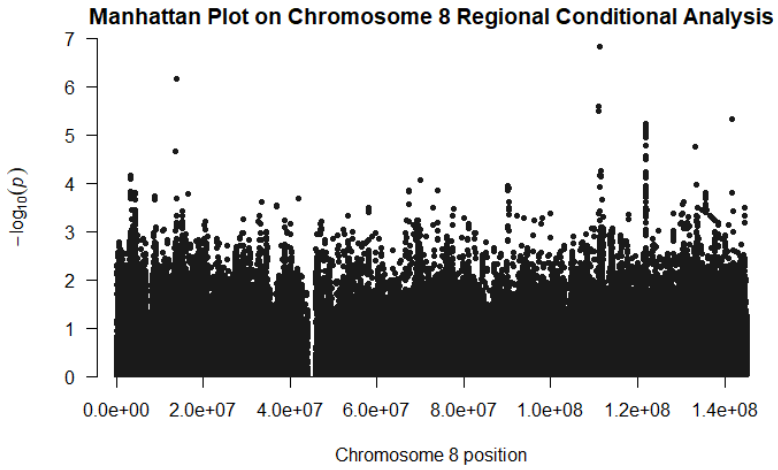


Figure 9: Manhattan Plot on Chromosome 8 Regional Conditional Analysis

We were not able to identify new signals on chromosome 8 conditioned on

rs1495741 from NAT2 gene.

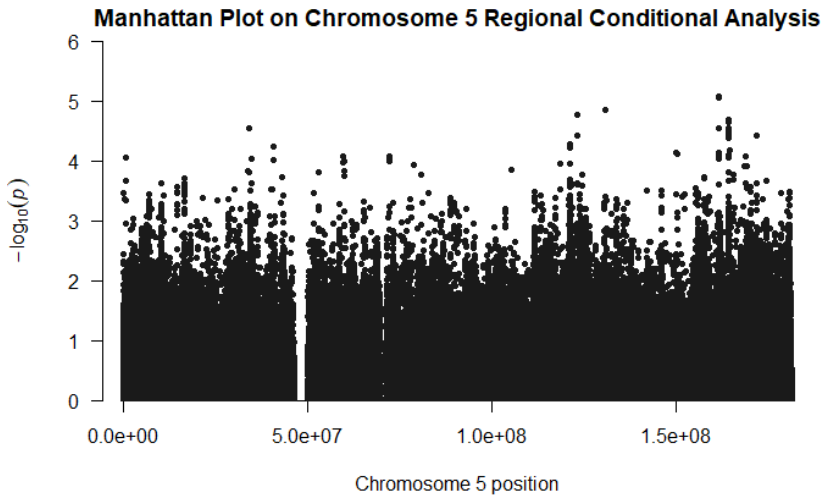


Figure 10: Manhattan Plot on Chromosome 8 Regional Conditional Analysis

We were not able to identify new signals on chromosome 5 conditioned on rs62342757.

Principal Component Analysis on SIFs

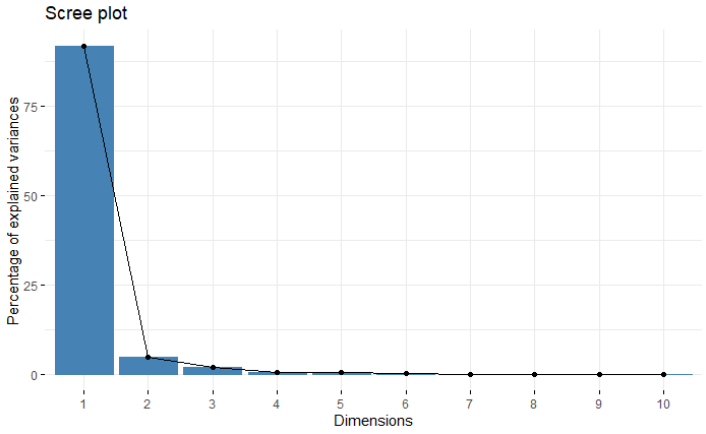


Figure 11: Scree Plot from SIFs Principal Component Analysis

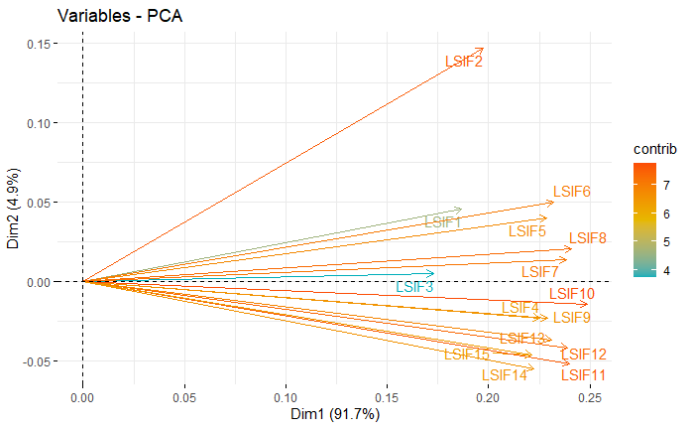


Figure 12: Variable Bi-Plot from SIFs Principal Component Analysis

LSIF: Logged SIF. Dim: Dimension

From Figure 11 & 12, first principal component of SIFs explains 91.7% of total variance. Replotted SIF vectors on Bi-plot are close to first principal component vector. This implies that SIFs are highly dependent, and their real dimension is low. Despite LSIF10 seems to be better choice for traits, LSIF1 was previously shown to be most correlated with AGEs in skin.

K-mean clustering of SIF and Covariates

Table 5: Clusters Table of K-mean

	Cluster 1	Cluster 2	Cluster 3
Genotype CC	135	101	186
Genotype CT	160	90	249
Genotype TT	50	32	79

An ideal result would be each cluster mainly having one genotype, but as we can see from Table 5, k-mean clusters are not able to distinguish genotypes.

Partition Around Medoids clustering of SIF and Covariates

Table 6: Clusters Table of PAM

	Cluster 1	Cluster 2	Cluster 3

Genotype CC	181	124	117
Genotype CT	247	150	102
Genotype TT	73	48	40

Like k-mean, PAM clusters are not able to distinguish genotypes.

Discussion

Our GWAS has shown that PR domain zinc finger protein 9 gene influences skin intrinsic fluorescence measure. *PRDM9* is known to be influencing meiotic recombination, with its zinc finger determining locations of recombination (**Paigen, 2018, 13**).

Two previous GWAS on SIF1 identified NAT2 gene and rs7533564 (**Eny, 2014, 3**) (**Roshandel, 2016, 4**). We were able to detect NAT2 gene signals but not rs7533564 in this study. This could be due to difference in data and model. GWAS detected rs7533564 used combined data from DCCT and Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR), also NAT2 gene was included as covariates in the regression model.

GWAS finding shows potential relationships between meiotic recombination and build up of advanced glycation end-product in skin. However, p-value was marginally significant ($3.9 * 10^{-8} < 5 * 10^{-8}$), there were only two SNPs from *PRDM9* regulating region detect rather than a large cluster of SNPs like NAT2, so a replication on different dataset is required.

Detecting *PRDM9* effect can help us better understand long term glycation build up in T1D patients with European descent. Thus, giving insights to researchers who works on how to reduce long term damage from T1D.

Acknowledgement

The authors acknowledge the patients and researchers of DCCT/EDIC.

Special thanks to Prof. Andrew Paterson and Dr. Delnaz Roshandel.

Reference

1. "Diabetes in Canada." , Public Health Agency of Canada, 14 Nov. 2017, www.canada.ca/en/public-health/services/publications/diseases-conditions/diabetes-canada-highlights-chronic-disease-surveillance-system.html#box1.
2. Cade, W Todd. "Diabetes-related microvascular and macrovascular diseases in the physical therapy setting." Physical therapy vol. 88,11 (2008): 1322-35. doi:10.2522/ptj.20080008
3. Eny, K.M., Lutgers, H.L., Maynard, J. *et al.* GWAS identifies an *NAT2* acetylator status tag single nucleotide polymorphism to be a major locus for skin fluorescence. *Diabetologia* **57,** 1623–1634 (2014). doi:10.1007/s00125-014-3286-9
4. Roshandel D, J. *et al.* New Locus for Skin Intrinsic Fluorescence in Type 1 Diabetes Also Associated With Blood and Skin Glycated Proteins. Diabetes. 2016 Jul;65(7):2060-71. doi: 10.2337/db15-1484. Epub 2016 Apr 12. PMID: 27207532; PMCID: PMC4915582.
5. Taliun, D. et al. (2019) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Biorxiv, doi:10.1101/563866
6. Diabetes Control and Complications Trial Research Group et al. "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus." The New England journal of medicine vol. 329,14 (1993): 977-86. doi:10.1056/NEJM199309303291401
7. Cleary, Patricia A et al. "Clinical and technical factors associated with skin intrinsic fluorescence in subjects with type 1 diabetes from the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study." Diabetes technology & therapeutics vol. 15,6 (2013): 466-74. doi:10.1089/dia.2012.0316
8. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P.-R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nature Genetics, 48(10), 1284–1287.
9. Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2014). minimac2: faster genotype imputation. Bioinformatics, 31(5), 782–784.
10. Dudbridge, Frank, and Arief Gusnanto. "Estimation of significance thresholds for genomewide association scans." Genetic epidemiology vol. 32,3 (2008): 227-34. doi:10.1002/gepi.20297
11. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. biorXiv DOI: 10.1101/005165 (2014).
12. Pruim, Randall J et al. "LocusZoom: regional visualization of genome-wide association scan results." Bioinformatics (Oxford, England) vol. 26,18 (2010): 2336-7. doi:10.1093/bioinformatics/btq419

13. Paigen K, Petkov PM. PRDM9 and Its Role in Genetic Recombination. Trends Genet. 2018 Apr;34(4):291-300. doi: 10.1016/j.tig.2017.12.017. Epub 2018 Jan 21. PMID: 29366606; PMCID: PMC5878713.

Appendix

Imputed genetics data was given in the beginning, so not able to provide code regarding TOPMed imputation.

Locus zoom plot was produced using (https://my.locuszoom.org/).

SNPTEST code

GWAS

Loop over 22 chromosomes

```
#!/bin/bash -x

#PBS -l walltime=24:00:00

cd /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1

let a=1

while [ $a -le 22 ]
do
        qsub    -v PARAM1=$a Analysis2.sh
      let a=$a+1
done
```

Code for each chromosome

```
#$ -S /bin/bash

#PBS -l vmem=30g,mem=30g
#PBS -l walltime=24:00:00

#PBS -o /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/queue
#PBS -e /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/queue

#change to current work directory
cd /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1
```

```
chr=$PARAM1

module load snptest/2.5

snptest \
-data
/hpf/largeprojects/andrew/adp/Delnaz/DCCT_EDIC_Ill1M_TopMed_Imputation/McCarthy_Freq_Diff_Off/TopMed_Freq_Fil
tre_Off/chr${chr}.dose.vcf.gz \
/hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/Pheno_Covars.txt \
-genotype_field GP \
-frequentist 1 \
-method expected \
-use_raw_phenotypes \
-pheno LSIF1 \
-cov_names Age Sex smoke_year_1_5 smoke_year_6_10 smoke_year_11_16 MREFSUM NORSOUTH MEANGFR
    Time_Weighted_HbA1c \
-o /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/out/DCCT_chr${chr}.res \
-log /hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/log/DCCT_chr${chr}.log

awk 'FNR>15 && $21!="NA" && $9>=0.8 && $19>=0.01 {print
"""$chr""",$2,$4,$5,$6,$9,$10,$11,$12,$18,$19,$21,$23,$24}'
/hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/out/DCCT_chr${chr}.res >
/hpf/largeprojects/andrew/adp/Zeya/dcct_gwas_topmed_sif1/temp/DCCT_chr${chr}_MAF_INFO_Checked.tmp
```

Regional Conditional Analysis

Chromosome 5

```
#PBS -l vmem=30g,mem=30g
#PBS -l walltime=24:00:00

#PBS -o /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/queue
#PBS -e /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/queue

#change to current work directory
cd /hpf/largeprojects/andrew/adp/Zeya/dcct_regional

module load snptest/2.5

snptest \
-data
/hpf/largeprojects/andrew/adp/Delnaz/DCCT_EDIC_Ill1M_TopMed_Imputation/McCarthy_Freq_Diff_Off/TopMed_Freq_Fil
```

```
tre_Off/chr5.dose.vcf.gz \
/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/Pheno_Covars1.txt \
-genotype_field GP \
-frequentist 1 \
-method expected \
-use_raw_phenotypes \
-pheno LSIF1 \
-cov_names Age Sex smoke_year_1_5 smoke_year_6_10 smoke_year_11_16 MREFSUM NORSOUTH MEANGFR
    Time_Weighted_HbA1c CHR5SNP \
-o /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/out/DCCT_chr5_1snp.res \
-log /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/log/DCCT_chr5_1snp.log

awk 'FNR>15 && $21!="NA" && $9>=0.8 && $19>=0.01 {print $2,$4,$5,$6,$9,$10,$11,$12,$18,$19,$21,$23,$24}'
/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/out/DCCT_chr5_1snp.res >
/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/temp/DCCT_chr5_1snp_MAF_INFO_Checked.tmp
```

Chromosome 8

```
#PBS -l vmem=30g,mem=30g
#PBS -l walltime=24:00:00

#PBS -o /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/queue
#PBS -e /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/queue

#change to current work directory
cd /hpf/largeprojects/andrew/adp/Zeya/dcct_regional

module load snptest/2.5

snptest \
-data
/hpf/largeprojects/andrew/adp/Delnaz/DCCT_EDIC_Ill1M_TopMed_Imputation/McCarthy_Freq_Diff_Off/TopMed_Freq_Fil
tre_Off/chr8.dose.vcf.gz \
/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/Pheno_Covars1.txt \
-genotype_field GP \
-frequentist 1 \
-method expected \
-use_raw_phenotypes \
-pheno LSIF1 \
-cov_names Age Sex smoke_year_1_5 smoke_year_6_10 smoke_year_11_16 MREFSUM NORSOUTH MEANGFR
    Time_Weighted_HbA1c CHR8SNP \
-o /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/out/DCCT_chr8_1snp.res \
-log /hpf/largeprojects/andrew/adp/Zeya/dcct_regional/log/DCCT_chr8_1snp.log

awk 'FNR>15 && $21!="NA" && $9>=0.8 && $19>=0.01 {print $2,$4,$5,$6,$9,$10,$11,$12,$18,$19,$21,$23,$24}'
```

/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/out/DCCT_chr8_1snp.res >
/hpf/largeprojects/andrew/adp/Zeya/dcct_regional/temp/DCCT_chr8_1snp_MAF_INFO_Checked.tmp

R code

```
## Data merging, pre-processing
library("readxl")
df1 <- read.csv("Covariates_LSIFs_1082.csv")
df2 <- read_excel("Order.xlsx")

colnames(df2) <- c("Order","ID_2","FID","TopMed_ID")
df3 <- merge(x = df2, y = df1[-1,],by = "ID_2", all.x = TRUE)
df3 <- df3[order(df3$Order),]
df4 <- df3[,-c(1:5)]
df4$ID_1 = df3$TopMed_ID
df4$ID_2 = df3$TopMed_ID

df <- rbind(df1[1,-14],df4[,c(12,13,1:11)])
df$missing[is.na(df$missing)] <- 0
sum(is.na(df$missing))
sum(is.na(df$Age))

write.table(df,file = "Pheno_Covars.txt", sep = "    ", dec = ".",
                row.names = FALSE, col.names = TRUE, quote = FALSE)

library(haven)
df2 <- read_excel("Mean_DCCT_Caffeine.xlsx")
dft <- dft[order(dft$FID),]
dft <- merge(x = dft, y = df2,by = "FID", all.x = TRUE)
df1 <- read_sas("ages.sas7bdat")
df1 <- na.omit(df1)
dfs <- dft[,c(1,15)]
dfa <- merge(x = df1, y = dfs,by = "FID", all.x = TRUE)

write.csv(dfa,file = "agesdat.csv",row.names = FALSE)

df1 <- read_sas("ga_expt.sas7bdat")
df2 <- read_sas("longbgp1.sas7bdat")

dft$chr5sigh <- ifelse(dft$chr5sig < 0.5,0,ifelse(dft$chr5sig < 1.5,1,2))

write.csv(dft,file = "fulldat.csv",row.names = FALSE)

df2 <- read.csv("fulldat.csv")
```

```
df3 <- df2[,c(1,13,15)]
names(df3) <- c("PATIENT","LSIF1","chr5sig")

df1[,c(1:10,20:89)] <- log(df1[,c(1:10,20:89)])
dfi <- df1[df1$`_Imputation_`==1,]
df1$mga <- rowMeans(df1[,c(1:10)],na.rm = TRUE)
df1$mbg <- rowMeans(df1[,c(20:89)],na.rm = TRUE)
df1 <- df1[,c("PATIENT","_Imputation_","GROUP","CVD","RET","REN","mga","mbg")]
dfi1 <- df1[df1$`_Imputation_`==1,]
dfi2 <- df1[df1$`_Imputation_`==2,]
dfi3 <- df1[df1$`_Imputation_`==3,]
dfi4 <- df1[df1$`_Imputation_`==4,]
dfi5 <- df1[df1$`_Imputation_`==5,]
dfi6 <- df1[df1$`_Imputation_`==6,]
dfi7 <- df1[df1$`_Imputation_`==7,]
dfi8 <- df1[df1$`_Imputation_`==8,]
dfi9 <- df1[df1$`_Imputation_`==9,]
dfi10 <- df1[df1$`_Imputation_`==10,]

dfc1 <- na.omit(merge(dfi1,df3,by = "PATIENT",all.x = TRUE))
dfc2 <- na.omit(merge(dfi2,df3,by = "PATIENT",all.x = TRUE))
dfc3 <- na.omit(merge(dfi3,df3,by = "PATIENT",all.x = TRUE))
dfc4 <- na.omit(merge(dfi4,df3,by = "PATIENT",all.x = TRUE))
dfc5 <- na.omit(merge(dfi5,df3,by = "PATIENT",all.x = TRUE))
dfc6 <- na.omit(merge(dfi6,df3,by = "PATIENT",all.x = TRUE))
dfc7 <- na.omit(merge(dfi7,df3,by = "PATIENT",all.x = TRUE))
dfc8 <- na.omit(merge(dfi8,df3,by = "PATIENT",all.x = TRUE))
dfc9 <- na.omit(merge(dfi9,df3,by = "PATIENT",all.x = TRUE))
dfc10 <- na.omit(merge(dfi10,df3,by = "PATIENT",all.x = TRUE))

pc <- read.delim("snp covariate model result/Pheno_Covars.txt", sep = "")
pc1 <- pc
pc1$LSIF1 <- ifelse(pc$Sex==2,NA,pc$LSIF1)
pc2 <- pc
pc2$LSIF1 <- ifelse(pc$Sex==1,NA,pc$LSIF1)

write.table(pc1,file = "xchr/Pheno_Covars1.txt", sep = "    ", dec = ".",
            row.names = FALSE, col.names = TRUE, quote = FALSE)
write.table(pc2,file = "xchr/Pheno_Covars2.txt", sep = "    ", dec = ".",
            row.names = FALSE, col.names = TRUE, quote = FALSE)

## Histogram of SIF1 and LSIF1, Descriptive Statistics

hist(dft$exp(LSIF1))
```

```
hist(dft$LSIF1)
summary(dft)

## Top 50 Hits from GWAS

genome <- read.delim("pval/DCCT_Chr1_22_With_Header.txt",sep = " ")
genomerank <- genome[order(genome$P),]
genomerank[1:50,]

## Manhattan Plot

library(qqman)
manhattan(genome,suggestiveline = F, main = "Manhattan Plot of SNPs")

## Histogram and QQ-Plot of P-Value

hist(genomerank$P, freq = F, main = "Histogram of P-value",xlab = "P-value")
abline(h=1,lwd =2,col ="blue")
qq(genomerank$P, main = "Quantile-Quantile Plot of P-value")

## GC lambda Calculation

gclambda <- function(gwas) {

    z <- qnorm(gwas
    $P / 2)

    lambda <- median(z^2) / qchisq(0.5, 1)

    return(lambda)}

gclambda(genome)

## Stratified Density Plot

dft$chr5sigh <- ifelse(dft$chr5sig > 1.5, 2, ifelse(dft$chr5sig > 0.5, 1, 0))
library(tidyverse)
theme_set(theme_bw(16))
source("https://raw.githubusercontent.com/datavizpyr/data/master/half_flat_violinplot.R")
library(ggplot2)
meansnp <- c(mean(dft$LSIF1[dft$chr5sigh==0]),mean(dft$LSIF1[dft$chr5sigh==1]),mean(dft$LSIF1[dft$chr5sigh==2]))
means <- round(meansnp,3)
rs62342757 <- as.factor(dft$chr5sigh)
rs62342757 <- ifelse(rs62342757==2,"TT, n = 161",ifelse(rs62342757==0,"CC, n = 422","CT, n = 499"))
```

```
ggplot(dft, aes(rs62342757,LSIF1,fill=rs62342757)) +
    geom_flat_violin() +
    geom_segment(aes(x=1,y=meansnp[1],xend=2,yend=meansnp[2]), colour="black") +
    geom_segment(aes(x=2,y=meansnp[2],xend=3,yend=meansnp[3]), colour="black") +
    geom_text(label = means[1], aes(x=0.9,y=meansnp[1]+0.1)) +
    geom_text(label = means[2], aes(x=1.9,y=meansnp[2]+0.1)) +
    geom_text(label = means[3], aes(x=2.9,y=meansnp[3]+0.1)) +
    theme(legend.position="none") +
    labs(title = "rs62342757 Genotype Stratified LSIF1 Density Plot",x = "rs62342757 Genotypes")

## Regression Model with other SIF Measures

m15 <- lm(LSIF2 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m15)

m16 <- lm(LSIF3 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m16)

m17 <- lm(LSIF4 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m17)

m18 <- lm(LSIF5 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m18)

m19 <- lm(LSIF6 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m19)

m20 <- lm(LSIF7 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m20)

m21 <- lm(LSIF8 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m21)

m22 <- lm(LSIF9 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m22)
```

```
m23 <- lm(LSIF10 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m23)

m24 <- lm(LSIF11 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m24)

m25 <- lm(LSIF12 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m25)

m26 <- lm(LSIF13 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m26)

m27 <- lm(LSIF14 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m27)

m28 <- lm(LSIF15 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH
+ MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dft)
summary(m28)

## Test rs62342757 with covariates

m6 <- lm(MREFSUM ~ chr5sig,data = dfc)
summary(m6)

m8 <- lm(MEANGFR ~ chr5sig,data = dfc)
summary(m8)

m10 <- lm(Time_Weighted_HbA1c ~ chr5sig,data = dfc)
summary(m10)

## Residual Plot of rs62342757 regression

m1 <- lm(LSIF1 ~ Age + Sex + smoke_year_1_5 + smoke_year_6_10 + smoke_year_11_16 + MREFSUM + NORSOUTH +
MEANGFR + Time_Weighted_HbA1c + chr5sig,data = dfc)
summary(m1)
par(mfrow=c(2,2))
plot(m1)

## Regional Conditional Analysis and Plotting
```

```
regional1snp <- read.delim("pval/DCCT_chr5_1snp_MAF_INFO_Checked.tmp",sep = " ",header = FALSE)
colnames(regional1snp) <-
c("rsid","BP","alleleA","alleleB","info","cohortAA","cohortAB","cohortBB","total","maf","P","beta","se")
regional1snprank <- regional1snp[order(regional1snp$P),]
regional1snprank[1:50,]
regional1snp$CHR = 5
manhattan(regional1snp,suggestiveline = F, main = "Manhattan Plot on Chromosome 5 Regional Conditional Analysis")


regionalchr81snp <- read.delim("pval/DCCT_chr8_1snp_MAF_INFO_Checked.tmp",sep = " ",header = FALSE)
colnames(regionalchr81snp) <-
c("rsid","BP","alleleA","alleleB","info","cohortAA","cohortAB","cohortBB","total","maf","P","beta","se")
regionalchr81snprank <- regionalchr81snp[order(regionalchr81snp$P),]
regionalchr81snprank[1:50,]
regionalchr81snp$CHR = 8
manhattan(regionalchr81snp,suggestiveline = F, main = "Manhattan Plot on Chromosome 8 Regional Conditional Analysis")

## Principal Component Analysis on SIFs

library(factoextra)
dfs <- dft[,c(12,15:28)]
dp <- prcomp(dfs)
par(mfrow=c(2,1))
fviz_eig(dp)
fviz_pca_var(dp,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )

## K-mean

km.out <- kmeans(dft[,c(3,5,6,7,8,10,11,12)],3,nstart=50)
table(dft$chr5sigh,km.out$cluster)

## PAM

library(cluster)
diss <- daisy(dft[,c(3,5,6,7,8,10,11,12)])
pam.out <- pam(diss, k=3)
table(dft$chr5sigh,pam.out$cluster)
```

End of Report