HIGH SCHOOL EDITION
Journal of Student Research
Volume 10 Issue 3 (2021)

Examining the Influence of Genes Linked to Comorbidities on the Development of Diabetic Retinopathy

Ishani Das[1] and Larry DeMuth[#]

[1]Cupertino High School, Cupertino, CA, USA
[#]Advisor

ABSTRACT

Diabetic retinopathy (DR) is a complication in diabetic patients that can cause vision loss and even blindness[12] {Healthline}. The condition causes bleeding in the blood vessels of the retina and ultimately interferes with the capture of photons and the transmission of neural signals to the brain[3] {National Eye Institute}. Unfortunately, there is no cure available for the approximately 200,000 US citizens currently affected by DR[8] {Lucas Research}. Current treatments are merely palliative and rarely restore vision. The best approach to combating DR is prevention. Understanding the root genetic causal links of DR is crucial to developing effective prevention plans. Unlike Huntington's Disease, which has well defined genetic origins, no single gene has been pinpointed as a causative factor for DR, despite having been extensively researched. Since diabetes alone isn't enough to predict the progression to DR[4]{NCBI}, one can hypothesize that a combination of diabetes and at least one other comorbidity, such as high blood pressure, high cholesterol, etc., may be essential to the development of DR. In this way, single nucleotide polymorphisms (SNPs) associated with other comorbidities could increase the likelihood of retinal bleeding in diabetic populations. The identification of these SNPs would offer valuable insight into the development of a genetically-driven prevention plan for diabetic retinopathy.

Introduction

In this work, the relationship between genes linked to common comorbidities and their apparent implications in the progression of DR were explored. This was examined by comparing genome-wide association studies (GWAS) for DR with that of other comorbidities, such as high blood pressure and high cholesterol. GWAS is an approach to reveal how strongly certain genetic variations, such as SNPs, are correlated with a particular disease. Scanning the entire genome of a person (all the DNA and genetic material) allows researchers to predict the presence of different diseases. One can hypothesize that if the SNPs related to other underlying conditions are also common to diabetic retinopathy, then it can be determined that DR is not directly caused by a single specific gene, but rather by a host of genes that increase the tendency of a person to develop it, e.g. a gene linked to high blood pressure does not directly cause DR itself, but it might increase the probability of a patient with diabetes developing DR. Lastly, this research investigated whether the SNPs underlying other common conditions are also associated with the progression of diabetic retinopathy, with the hope of developing effective prevention plans for patients with diabetes in the future.

To investigate this relationship, the SNPs linked to other comorbidities were examined and studied. From this research, it was discovered that the gene TULP4 (Tubby-related protein 4) may be correlated to the development of diabetic retinopathy. 4 out of the 6 total SNPs shared common between diabetic retinopathy and diastolic blood pressure with a p-value of less than 0.0001, were either mutations in or directly upstream of the TULP4 gene (Table 1).

This gene codes for the protein Tubby-related protein 4, which is involved in the pathway of protein ubiquitination, which helps regulate the processes of other proteins in the body. Therefore, it is possible that mutations in the TULP4 gene cause a comorbidity that leads to the development of DR in diabetic patients. Being just one of many comorbidities, the work presented here suggests that a more extensive study could reveal much deeper and predictive connections between SNPs and the development of DR.

Methods

Data Collection

First, reliable datasets containing the necessary GWAS information for several diseases were collected. This included SNPs (single-nucleotide polymorphisms) that could be causative factors for the specific disease over the full human genome. An SNP is a substitution of a single nucleotide at a specific position in the genome. These nucleotides are what build DNA (deoxyribonucleic acid) which code for proteins. Misfolding or truncation of these proteins can give rise to many diseases and health complications. Figure 1 shows an example of an SNP where the nucleotide C (cytosine) is switched to A (adenine).

This research utilized the database "Type 2 Diabetes Knowledge Portal" [9] {T2D}, which provided the information being searched for - datasets containing the SNP information for a variety of phenotypes including ocular, cardiovascular, and lipids. From there, the data for Diabetic Retinopathy[10] {NCBI}, HDL/LDL/Total Cholesterol[6] {Center for Statistical Genetics}, Triglycerides[6] {Center for Statistical Genetics}, Heart Failure[11] {NCBI}, Age-Related Macular Degeneration[1] {Center for Statistical Genetics}, and Systolic/Diastolic Blood Pressure[5] {NCBI} were downloaded from reliable sites, all partnered with the government agency, NIH (National Institute of Health) through the "Type 2 Diabetes Portal". Whenever possible, the research ensured that the GWAS datasets were derived were from large, ethnically diverse populations (greater than 10,000 participants from a mix of ethnicities).

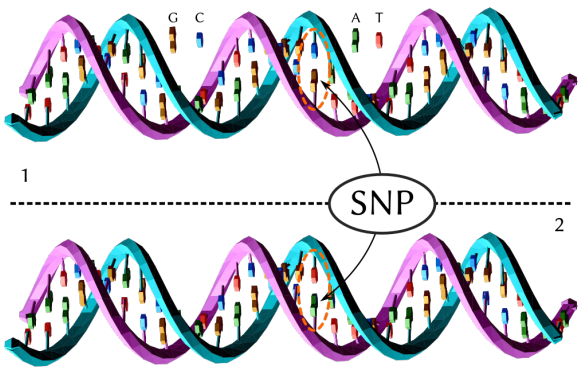

Figure 1. | Example of an SNP[14] {Image reproduced from Wikimedia Commons. Created by David Eccles. License under 4.0 International.} This picture is a diagram of an SNP (single-nucleotide polymorphism). These SNPs (also known as rsIDs) are what were used as the method of identifying which comorbidities were must correlated to diabetic retinopathy.

Data Collection

After collecting the data, custom code was written to parse these datasets and scrape out the information necessary to reach a conclusion - the rsID (name for each unique SNP), and p-value, a statistical parameter that represents how strongly correlated that SNP is to the disease or condition in question. Python written in Jupyter notebook was used to create a program that parsed out the rsIDs and p-value given a certain threshold. By doing this, analysis was able

to focus on only a certain portion of the SNPs. These diseases were then compared to diabetic retinopathy, and graphed the percentage of shared SNPs over a range of five p-values (0.0001, 0.001, 0.01, 0.1, 1).

Data Analysis & Interpretation

To figure out which diseases were most correlated to diabetic retinopathy, the last set of rsIDs in the lowest p-value of each set was identified and ensured to not be empty. From there, the background information necessary to study the biological background and find the common factor that caused or resulted from the certain SNPs was researched with information from the NCBI (National Center for Biotechnology Information), GeneCards[2], and UniProt[13]. From this research, it was also discovered that the diseases that had a higher correlation from Figure 2, also had SNPs that were closer in terms of the chromosome position than those which were less correlated, which will be elaborated more on in the "*Results*" section.

Results

The study resulted in the determination that Diastolic Blood Pressure Blood Pressure, LDL Cholesterol, and Total Cholesterol had the highest correlation to diabetic retinopathy. This was determined by comparing the percentage of SNPs shared between diabetic retinopathy with other diseases as a function of p-value. When plotting the log (p-value) vs. log (percentage), it can be seen that the three aforementioned conditions share a much higher fraction of SNPs even at very low p-values (Figure 2). For example, ~0.01% of SNPs are shared between diabetic retinopathy and LDL Cholesterol at a p-value of 10^{-4}, which is much higher than that of the other comorbidities, such as Systolic Blood Pressure and HDL Cholesterol, as they are around ~0.001% at the same p-value of 10^{-4}. The remaining comorbidities were even lower. Taking into account that all GWAS data was a result of scanning the entire genome of several heterogeneous populations, and the diabetic retinopathy contained over 8 million SNPs total, having 0.01% common genes is very statistically significant. With the data collected from the findings, further research was performed into studying the background of each of the rsIDs at the lowest p-values for the highly correlated diseases and collected the following data: chromosome position, alleles, gene, gene significance, variant, and frequency in the population, from the NCBI website. Table 1 shows an example of this information for all the common SNPs shared between Diastolic Blood Pressure and diabetic retinopathy at the p-value threshold of 0.0001. The same data for LDL and Total Cholesterol are presented in Table 2.

After comparing the three most highly-correlated comorbidities (diastolic blood pressure, LDL cholesterol Total cholesterol) with each other, it was found that diastolic blood pressure and total cholesterol shared the most common SNPs with each other a at a p-value of log(-3). However, LDL cholesterol and Total cholesterol were expected to share the most common SNPs at the lowest non-zero p-value, as all five SNPs common between DR and LDL cholesterol, as well as DR and total cholesterol at a p-value of 0.0001 were the same.

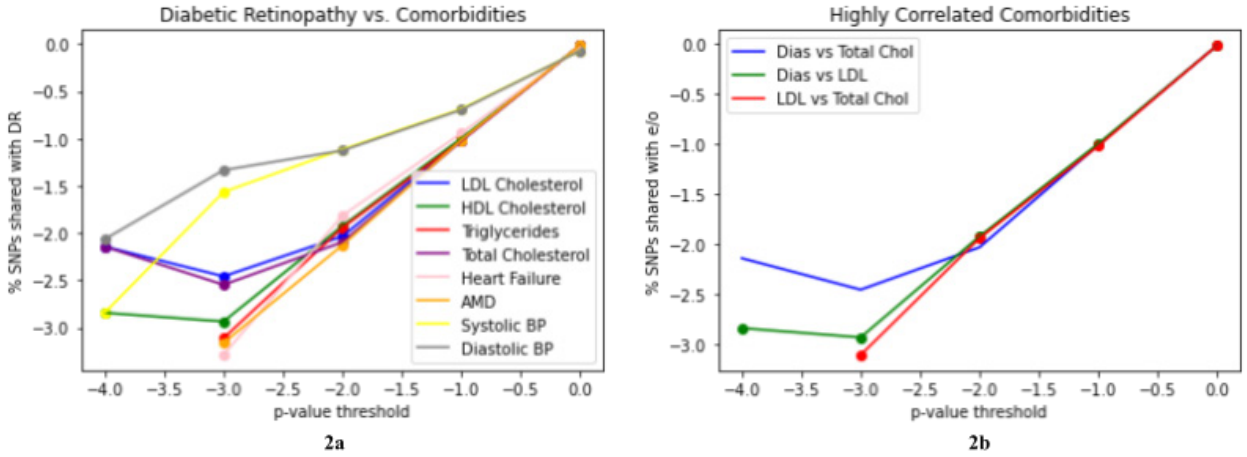

Figure 2. | Trial results. a) The comparison of fraction of SNPs shared between DR and eight common comorbidities (LDL/HDL/Total Cholesterol, Triglycerides, Heart Failure, age-related macular degeneration, and systolic/diastolic blood pressure) shows the percent of common SNPs between the comorbidities and DR over the following comorbidities: [0.0001, 0.001, 0.01, 0.1, 1], which can be used to determine which comorbidities are most correlated to DR at the lowest non-empty SNP dataset p-value. b) Isolating the top three most correlated comorbidities (diastolic blood pressure and LDL/total cholesterol) with each other showed was similarly used to determine which of those comorbidities are most correlated to each other at the lowest non-empty SNP dataset p-value.

Table 1 | DR + Diastolic BP Common rsIDs (p-value: 0.0001). The table contains information regarding the Position, Alleles, Variation Type, Gene + Consequence, Gene Significance, and Frequency for each SNP common between Diastolic Blood Pressure and DR at a p-value of 0.0001, the lowest non-empty dataset of common SNPs.

RSID	Position	Alleles	Variation Type	Gene + Consequence	Gene Significance	Frequency
rs11695443	chr2:158449810 (GRCh38.p12)	T → G	SNV	-CCDC148 -Intron Variant	-Protein: Coiled-Coil Domain Containing -Helps wrap/up-wrap DNA -Associated with Vulva Cancer and Vulvar Disease	7.6%
rs223104	chr3:169382119 (GRCh38.p12)	T →C	SNV	-MECOM -Intron Variant	-Protein: Histone-lysine N-methyltransferase MECOM -Transcriptional regulator (binds and regulates DNA sequences at the promoter region) -Associated with CML (chronic myelogenous leukemia)	17%

rs4710077	chr6:158228235 (GRCh38.p12)	G → A	SNV	None	-Less than 5000 base pairs upstream of the gene TULP4	40%
rs62437125	chr6:158232799 (GRCh38.p12)	A →G	SNV	-TULP4 -Intron Variant	-Protein: Tubby-related protein 4 -Involved in protein ubiquitination (part of modification)	41%
rs6906984	chr6:158229369 (GRCh38.p12)	G →A G →C G →T	SNV	None	-Less than 3500 base pairs upstream of the gene TULP4	40%
rs6929952	chr6:158230363 (GRCh38.p12)	C →T	SNV	-TULP4:2KB -Upstream Variant	Less than 2500 base pairs upstream of the gene TULP4	40%

Table 2 | DR + LDL Cholesterol Common rsIDs and DR + Total Cholesterol Common rsIDs. *(p-value threshold: 0.0001)* The table contains information regarding the Position, Alleles, Variation Type, Gene + Consequence, Gene Significance, and Frequency for each SNP common between Diastolic Blood Pressure and DR at a p-value of 0.0001, the lowest non-empty dataset of common SNPs.

RSID	Position	Alleles	Variation Type	Gene + Consequence	Gene Significance	Frequency
rs12496413	chr3:12290337 (GRCh38.p12)	C →T	SNV	-PPARG -Intron Variant	-Regulate transcription of various genes -Associated with obesity and diabetes -PPARG regulates fat storage cell differentiation)	25%
rs2920500	chr3:12281914 (GRCh38.p12)	G →A G →C	SNV	None	-Nearest coding gene: PPARG (<5,500 base pairs away) -May cause Ischemic Stroke	49%
rs2972164	chr3:12292917 (GRCh38.p12)	T →A T →C	SNV	-PPARG -Intron Variant	-Regulate transcription of various genes -Associated with obesity and diabetes	47%
rs56395424	chr3:12194223 (GRCh38.p12)	G →A	SNV	None	-A little over 96000 base pairs away from rs12496413 located on PPARG	22%

					-Nearest coding gene: TIMP4 (~35,000 base pairs away) -TIMP4 may be involved in hormonal regulation + endometrial tissue remodeling	
rs6775191	chr3:12230699 (GRCh38.p12)	G → A	SNV	None	-Nearest coding gene: PPARG (<60000 base pairs away) -Prevalence of 0.36% in Type 2 diabetes	27%

Discussion

In this work, it is reported a possible gene linked to the development of diabetic retinopathy. 4 of the 6 SNPs identified as common between diabetic retinopathy and diastolic blood pressure with a p-value of less than 0.0001 were located in, or directly upstream of, the gene TULP4. This gene encodes for the protein Tubby-related protein 4, which is involved in the pathway of protein ubiquitination, which helps regulate the processes of other proteins in the body. Interestingly, 3 of the 4--rs4710077, rs6906984, and rs6929952 are located upstream of the gene itself, and the fourth, rs62437125, is located in an intron of the gene itself within chromosome 6. This suggests that mutations that interfere with the transcription and expression of the gene are enough to cause a phenotypic change, without even altering the protein itself. The upstream region could potentially be associated with a promoter region that affects the gene expression levels of TULP4, i.e. how "observable" the phenotype is. The discovery of these 4 SNPs is very significant because in the human genome of over 3 billion base pairs, a distance of <3500 base pairs apart is extremely close. Interestingly, all 5 SNPs common between DR and LDL cholesterol as well as DR and total cholesterol at a p-value of 0.0001 were the same. To further investigate this interesting relationship, custom software was created and used to find the common SNPs between LDL cholesterol and total cholesterol. From this study, it was found that 6160 SNPs were shared between the two diseases at a p-value of 0.0001. After finding the total number of SNPs for LDL and total cholesterol at this p-value, it was discovered that ~84% of the SNPs in LDL cholesterol were shared between the two diseases, and ~61% for total cholesterol. While a concrete conclusion cannot be made about the implications of these findings, these results suggest that further research could help elucidate the correlation between these comorbidities.

Another notable trend was that many of the SNPs shared between the three most correlated diseases with DR (diastolic blood pressure, LDL cholesterol, and total cholesterol) had many SNPs (at a p-value of 0.0001) located on chromosome 3. The nearest coding gene for 3 of the SNPs common between DR and LDL/total cholesterol were PPARG. This gene (Peroxisome Proliferator Activated Receptor Gamma), is known to be implicated in the pathology of diabetes, obesity, and various other comorbidities. Further research in this area is needed to understand the exact role of PPARG and other genes like TIMP4 (~35,000 base pairs away from PPARG) in the development of DR.

Conclusion

Overall, it was concluded that the gene TULP4 may be correlated to the development of diabetic retinopathy. 4 out of the 6 total SNPs shared common between diabetic retinopathy and diastolic blood pressure ($p<0.0001$) were in or directly adjacent to this gene, so it is possible that this area affects the gene expression and potentially the ultimate development of diabetic retinopathy.

The data comparing LDL cholesterol and total cholesterol also showed an interesting trend that both datasets shared the exact same genes at the lowest p-value of 0.0001, all located on chromosome 3. Knowing that total cholesterol is a combination of both HDL and LDL cholesterol, further research can be performed to determine if that is the cause for this trend.

The insights gained from this work will be important to guiding future investigations into the development of DR. For example, investigating individual patients from all genders and ethnicities will give a more holistic picture of how these factors affect the development of diabetic retinopathy with the influence of other comorbidities.

The work presented here suggests that additional studies are needed in the field of diabetic retinopathy to not pinpoint a single gene, but rather identify a host of causative mutations, potentially linked to other comorbidities, to better understand this disease and create preventative measures. Future research will benefit from looking at individual patient data, rather than aggregated GWAS studies.

Limitations

Limitations to this study could include collecting data from different different sources and of mixed ethnicities. Focusing on a single ethnicity or individual patient data may provide more specific results.

Acknowledgements

The National Center for Biotechnology Information (NCBI) offered several research tools that helped aid the research including the dbSNP. Healthline[12], GeneCards[2], and UniProt[13], as well as other sites listed in *References* also offered background information regarding diabetic retinopathy and other comorbidities.

References

"AMD Gene Consortium Study of Age Related Macular Degeneration." *Center for Statistical Genetics*, csg.sph.umich.edu/abecasis/public/amdgene2012/.

Database, GeneCards Human Gene. "GeneCards®: The Human Gene Database." *GeneCards*, www.genecards.org/.

"Diabetic Retinopathy." *National Eye Institute*, U.S. Department of Health and Human Services, www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/diabetic-retinopathy#:~:text=Diabetic retinopathy is caused by,vessels all over the body.

Ehret, Georg B, and Mark J Caulfield. "Genes for Blood Pressure: an Opportunity to Understand Hypertension." *European Heart Journal*, Oxford University Press, Apr. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC3612776/.

Gazal S;Loh PR;Finucane HK;Ganna A;Schoech A;Sunyaev S;Price AL; "Functional Architecture of Low-Frequency Variants Highlights Strength of Negative Selection across Coding and Non-Coding Annotations." *Nature Genetics*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/30297966/.

"Global Lipids Genetics Consortium Results." *Center for Statistical Genetics*, csg.sph.umich.edu/willer/public/lipids2013/.

"Home - SNP - NCBI." *National Center for Biotechnology Information*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/snp/

"Learn How Diabetes Affects Eyesight." *Lucas Research*, 29 June 2017, lucasresearch.org/diabetes-affects-eyesight/.

Metabolic Disorders Knowledge Portal - Home, t2d.hugeamp.org/.

Pollack S;Igo RP;Jensen RA;Christiansen M;Li X;Cheng CY;Ng MCY;Smith AV;Rossin EJ;Segrè AV;Davoudi S;Tan GS;Chen YI;Kuo JZ;Dimitrov LM;Stanwyck LK;Meng W;Hosseini SM;Imamura M;Nousome D;Kim J;Hai Y;Jia Y;Ahn J;Leong A;Shah K;Park KH;Guo X;Ipp E;Taylor KD; "Multiethnic Genome-Wide Association Study of Diabetic Retinopathy Using Liability Threshold Modeling of Duration of Diabetes and Glycemic Control." *Diabetes*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/30487263/.

Shah S;Henry A;Roselli C;Lin H;Sveinbjörnsson G;Fatemifar G;Hedman ÅK;Wilk JB;Morley MP;Chaffin MD;Helgadottir A;Verweij N;Dehghan A;Almgren P;Andersson C;Aragam KG;Ärnlöv J;Backman JD;Biggs ML;Bloom HL;Brandimarto J;Brown MR;Buckbinder L;Carey DJ;Chasman. "Genome-Wide Association and Mendelian Randomisation Analysis Provide Insights into the Pathogenesis of Heart Failure." *Nature Communications*, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/31919418/.

Team, the Healthline Editorial. "Retina Function, Anatomy & Anatomy | Body Maps." *Healthline*, Healthline Media, 22 Jan. 2018, www.healthline.com/human-body-maps/retina.

UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. "UniProt Consortium." *UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics*, www.uniprot.org/.

"File:Dna-SNP.svg." *Wikimedia Commons*, commons.wikimedia.org/wiki/File:Dna-SNP.svg.