JOURNAL OF
EMERGING INVESTIGATORS

Article

Refinement of Single Nucleotide Polymorphisms of Atopic Dermatitis related Filaggrin through R packages

Aniket Naravane[1], Lauren Taylor[1]
[1] Lake Travis High School, Austin, Texas

SUMMARY

In the United States, there are currently 17.8 million affected by atopic dermatitis (AD), commonly known as eczema. It is characterized by itching and skin inflammation. AD patients are at higher risk for infections, depression, cancer, and suicide. Genetics, environment, and stress are some of the causes of the disease. With the rise of personalized medicine and the acceptance of gene-editing technologies, AD-related variations need to be identified for treatment. Genome-wide association studies (GWAS) have associated the Filaggrin (FLG) gene with AD but have not identified specific problematic single nucleotide polymorphisms (SNPs). This research aimed to refine known SNPs of FLG for gene editing technologies to establish a causal link between specific SNPs and the diseases and to target the polymorphisms. The research utilized R and its Bioconductor packages to refine data from the National Center for Biotechnology Information's (NCBI's) Variation Viewer. The algorithm filtered the dataset by coding regions and conserved domains. The algorithm also removed synonymous variations and treated non-synonymous, frameshift, and nonsense separately. The non-synonymous variations were refined and ordered by the BLOSUM62 substitution matrix. Overall, the analysis removed 96.65% of data, which was redundant or not the focus of the research and ordered the remaining relevant data by impact. The code for the project can also be repurposed as a tool for other diseases. The research can help solve GWAS's imprecise identification challenge. This research is the first step in providing the refined databases required for gene-editing treatment.

INTRODUCTION

There are currently 17.8 million atopic dermatitis (AD) patients in the US, with many having an increased risk of lower quality of life, cancer, depression, and suicide (1, 2). The disease places a burden on the caregivers and family of the patient as well. A 2004 study asserts that AD costs roughly $4.2 billion in the US, which has likely gone up (1, 3). AD has been associated with cancer due to the possible interference in the epithelial barrier function, allowing the penetration of carcinogens and viruses like human papilloma virus (4, 5). AD is a complex disease caused by many factors including environment, diet, mental state, and genes. Twin pair studies have shown that genetic inheritance is a major factor for AD (6). Genome-wide association studies (GWAS) have identified the Filaggrin gene (*FLG)* as highly correlated with AD (6). The largest GWAS to date in *FLG* have involved upward of 21,000 participants. Furthermore, previous studies have shown that loss-of-function single nucleotide polymorphisms (SNPs) in *FLG* lead to AD (6, 7). *FLG* encodes for Profilaggrin, which is dephosphorylated to produce multiple Filaggrin polypeptides. Filaggrin, a vital protein in the epidermis, balances pH, chaperones structural proteins, and supports moisturization.

The current GWAS results have narrowed down the source of genetic diseases to specific genetic areas that may play a role based on the correlation between disease and SNPs (8). Moreover, GWASs have a wider scope of the population than case studies, which allows for the conclusions to be extrapolated to the general population. SNPs are genetic variations that occur in more than 1% of the general population's genome (9). However, GWAS does not identify the specific SNPs that cause the disease (10). This problem is a roadblock for functional studies and possible therapeutic drugs. Gene-editing-based treatment would require the precise locations of significant polymorphisms, which current data do not provide. Testing every possible identified polymorphism in a wet laboratory is not practical. Therefore, it begs the question of how GWAS data, in this case for AD, can be filtered for SNPs that may impact the protein structure to expedite identifying biological pathways that cause diseases. This research's aim was to generate a tool to narrow down possibly causal SNPs to be verified in a laboratory. The study produced an R script using Bioconductor packages to filter SNPs in the *FLG* and order them by the possible severity in altering the protein structure and hence the function.

The algorithm focused on the *FLG* with the goal of expanding to other genes in the future. According to National Center for Biotechnology Information's (NCBI) Variation Viewer (Genome Version GRCh38), *FLG* is located in 1q21.3 (152,302,165–152,325,239 nt) and includes three exons (**Figure 1**). Exon 3 (12047 bp) is the largest of the three exons and is the most significant because it includes all coding domains, including the calcium ion binding region (11). The calcium ion binding region is responsible for the dephosphorylation of Profilaggrin, leading to its cleavage into the active peptides (11).

The algorithm needed to consider the various types of SNPs as not all impact protein function equally or even at all. Only variants that may impact the protein structure were used for analysis in this study. Although SNPs in introns play an important role in regulating gene expression, this study focused on the SNPs in the coding regions, as they could cause changes in protein structure and increase the risk of AD (12). The rationale behind selecting coding over non-

Figure 1: *FLG* gene diagram. Visual representation of the *FLG* with the 3 exon regions and Ca (2+) binding region. Figure is not drawn to scale.

coding regions is that the phenotypic outcome of variations in coding regions is much more predictable than those in non-coding regions. The nucleotide changes in the coding region are further categorized into frameshift, nonsense, nonsynonymous, and synonymous SNPs.

Frameshift SNPs, resulting from the deletion of DNA base pairs, change the reading frame of the transcript. This type of SNP changes the whole protein structure after the variation. Nonsense SNPs, resulting from single base pair substitutions, abruptly introduce stop codons leading to a truncated protein structure. The location of these two types of SNPs with respect to functional domains decides the magnitude of their impact on the protein structure. However, there are surveillance processes, such as nonsense mediated decay (NMD), that catch truncated or misread mRNA. Thus, these mutations do not necessarily lead to deformations in the protein structure.

Nonsynonymous SNPs arising from single base pair substitutions result in a single amino acid (AA) change. Previous research has proven that there are functional consequences of nonsynonymous SNPs in other genes, like *IL8*, a biomarker for AD (13, 14). Such SNPs have lower impact if the substituted AAs have similar properties than substituted AAs with different properties. This difference in impacts can be quantified using the BLOSUM 62 substitution matrix with lower scores indicating substitutions of AA with different properties resulting in possible changes to the protein structure. Synonymous SNPs do not change the final AA sequence and have no impact on the protein structure. Therefore, synonymous SNPs were filtered out by the algorithm.

The algorithm's validity was tested by checking if SNPs, positively identified by previous research to cause AD, were located in the final SNP dataset. The research is of significant help to identify causal genetic variations as it narrows down the number of SNP required to test for a causal link. The algorithm will aid gene editing by providing genetic targets for AD, and other diseases in the future.

RESULTS

The study began with SNPs in the *FLG* from the NCBI Variation Viewer. The raw dataset did not include any structural variants but did include indels, addition or deletion of base pairs. There was a total of 11,020 unique SNPs. However, SNPs in non-coding regions had multiple entries,

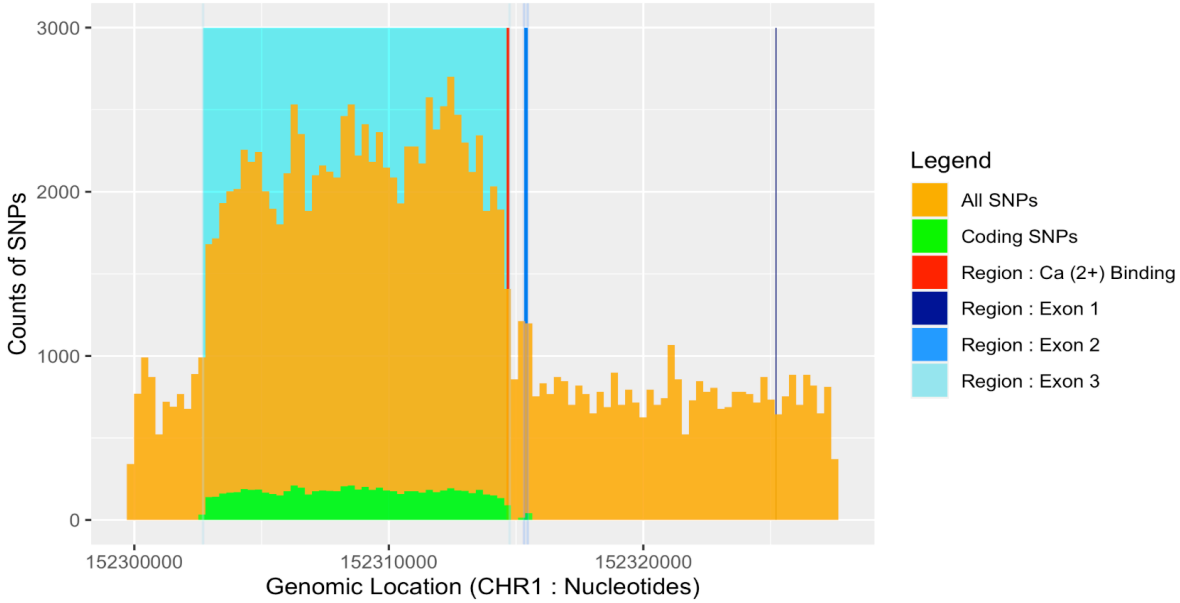

Figure 2: Pre-processed and semi-processed SNPs by genomic location. Total number of SNPs in the *FLG* genomic region (Orange) relative to SNPs within the coding regions (Green) for the genomic regions covering the three exons. Histogram based on the genomic location with markers of specific protein regions. The genomic regions of the exons are interspaced by numerous minor introns, which are not shown in the figure.

with each entry having slightly varied location information. The multiple entries increased the raw SNP dataset to a total of 129,323 entries. The first step was to remove all the SNPs in the non-coding region, which constituted a significant portion (94.18%) of the original data. The remaining variations (in green) are located primarily in exon 3, where all the *FLG* active domains are located, including the calcium-binding domain (in red, **Figure 2**). The remaining 5.8% (7,514 SNPs) of the original SNPs were sorted into nonsynonymous, synonymous, nonsense, and frameshift categories.

The R algorithm removed the synonymous SNPs, which accounted for most of the remaining SNPs (42.3% of SNPs in coding regions and 1,756 in total) because they do not result in an AA substitution and hence no change in the protein structure. The algorithm filtered out 1,418 nonsynonymous SNPs using the BLOSUM 62 scoring matrix from the original 5,082 SNPs (**Figure 3**). The remaining nonsynonymous SNPs were centered at about -2 and -1 on the BLOSUM 62 scale, where negative scores indicate amino acid changes that are not similar. There were 300 frameshift SNPs and 376 nonsense SNPs. In the end, 96.65% of the original data were removed, and the remaining relevant data were ordered based on impact on protein structure (**Figure 4**).

Previous research has specifically identified a nonsynonymous mutation (rs61816761/p.R501X) as a loss of function SNP strongly linked to AD (15). rs138726443, a

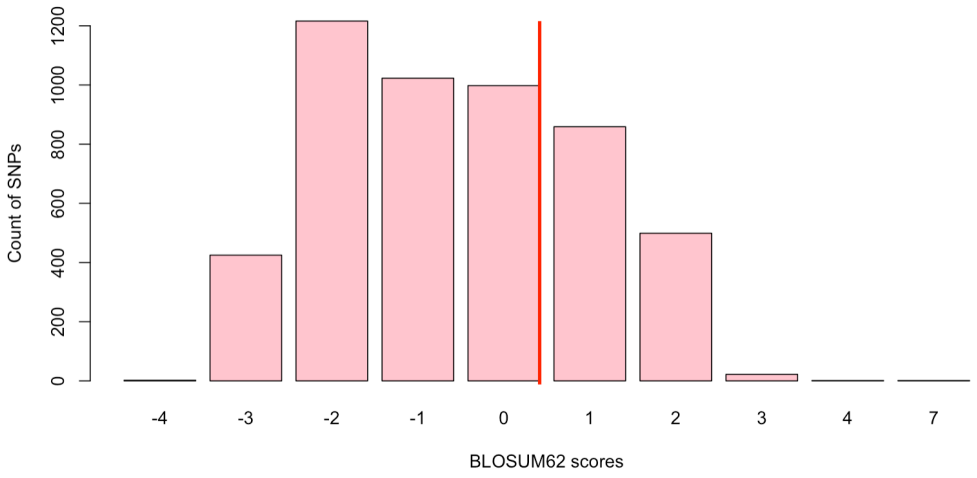

Figure 3: Distribution of BLOSUM62 scores. Lower scores indicate non-similar AA substitutions; the lower the number, the greater the dissimilarity from the native AA. Positive values indicate the same or AA with similar side chains and/or charge groups as the native protein. All SNPs that scored greater than 0 were not included in the data analysis. The cutoff is denoted by the red line.

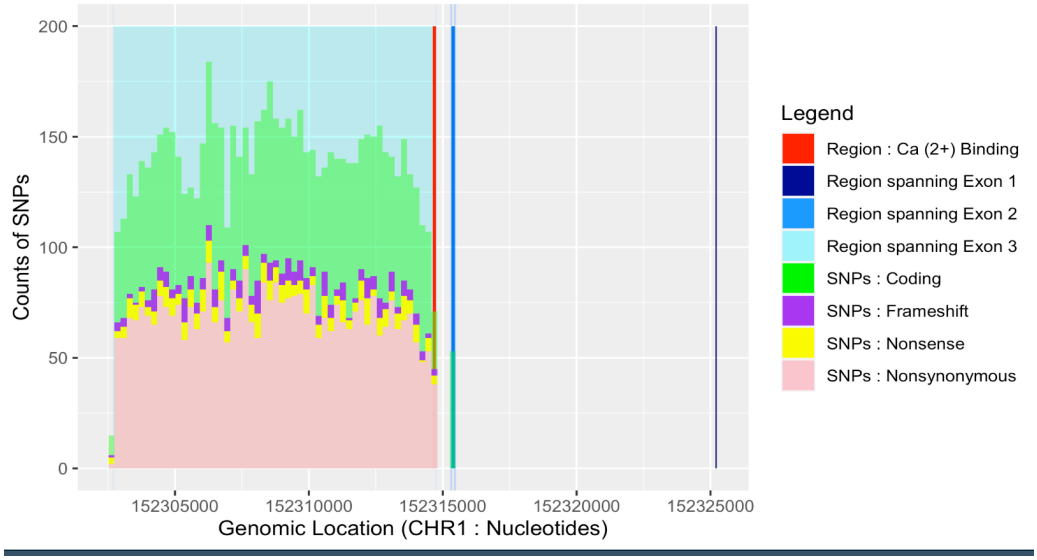

Figure 4: Final SNPs by Genomic Location. Stacked histogram of final frameshift (Purple), nonsynonymous (Pink), and nonsense (Yellow) SNPs with respect to all the coding SNPs (Green), genomic location, and markers of specific protein regions. The synonymous SNPs included in the total number of coding SNPs (Green) are not accounted for in the data analysis since they would not impact the protein structure. The genomic regions of the exons are interspaced by numerous minor introns, which are not shown in the figure.

nonsense mutation that results in a nonfunctional protein, also increases the risk for childhood AD (7). The presence of both SNPs in the final dataset confirms the validity of the selection strategy and the algorithm (**Figure 5**).

The R script identified 25 SNPs that may impact the Ca^{2+} binding region of the native protein because they could impact the production of the Filaggrin protein. These SNPs were tagged using an extra column in the final nonsynonymous mutation database with a binary system to indicate if the SNPs may impact the region.

DISCUSSION

In this study, the SNPs were filtered and ordered based on their impact on protein structure from the currently known SNPs in the NCBI's Variation Viewer database for the *FLG*. It is important to note that the data may not represent all the variations in the population.

The frameshift and nonsense SNPs were ordered based on genomic location because they significantly impact the protein structure when they occur earlier in the gene. While the algorithm for frameshift, nonsense, and synonymous SNPs was straightforward, the nonsynonymous SNPs required additional filtering. Nonsynonymous SNPs have varying impacts on protein structure because some SNPs interchange AAs with similar physicochemical properties and others with drastically different properties. The impact of the nonsynonymous polymorphisms on the protein structure was quantified using the BLOSUM62 substitution matrix. Positive scores indicate that the changed AA resembles the original AA, while negative scores indicate dissimilarity, so the polymorphisms with positive scores were removed. The distribution of the BLOSUM62 scores is skewed towards the negative because two different AAs cannot have similar properties, and extreme positive scores describe synonymous SNPs. Since extreme changes are negatively selected through evolution, the distribution is centered at -2 and -1 rather than -3 or less. This distribution represents the analysis of the data from the database and not in the population, which is a possible source of bias. After implementing the algorithm, 96.6% of the SNPs identified in the original database were removed to focus on 3.4% of the total SNPs to study their effect on AD further. The reduction makes the data manageable for conducting experimental studies to confirm the theoretical findings.

A possible way to improve the algorithm would be to integrate more GWAS data. Currently, the algorithm utilizes data from refSNPs of NCBI's Variation Viewer, which has a location and identification of the SNPs in the region but not GWAS data such as genome-wide significance threshold or Linkage Disequilibrium. By including more corresponding GWAS data, the algorithm will be strengthened and may prove to be more accurate.

Although the algorithm was successful in filtering and ordering the raw SNP dataset, as it positively identified the two previously known SNPs, more work needs to be done to validate the algorithm's accuracy further. Additional known SNPs, such as those in NCBI's ClinVar, need to be tested. However, since the *FLG* is not heavily studied, computational validation will be easier when the algorithm is modified for other common diseases. The integration of additional GWAS data could also test the accuracy if significant SNPs from GWAS are selected by the algorithm. Along with the computational aspect of the research, wet lab research needs to be conducted to support the project's validity further. A possible method would be using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) to introduce the SNP in cell cultures and comparing the impact of edited cultures to non-edited control cell cultures.

	rsID	pop. freq.	DNA variant	BLOSUM 62 *	Notes
SNP identified by Previous research	rs61816761	0.21517	G > A		
	rs138726443	0.00272	G > A		
Nonsynomous	rs370154185	0.00007	G > C	-3	Located in Ca(2+) binding site
	rs760681209	0.00003	A > G	-3	Located in Ca(2+) binding site
	rs780392363	0.000008	C > G	-3	
	rs150330138	0.000077	G > A	-3	
	rs200923584	0.000004	G > A	-3	
	rs752322228	0.00027	C > G	-2	
	rs746611119	0.00006	G > T	-2	
	rs142133815	0.00085	C > T	-2	
Nonsense	rs769442034	0.00005	C > A		
	rs771179731	0.000008	G > A		
	rs377051179	0.00007	G > A		
Frameshift	rs778511241	0.00941	A > G		
	s775583931	0.00004	C > A		
	rs1191126363	0.000008	G > A		

Figure 5: Table of select SNPs. A number of SNPs identified by previous research and in the nonsynonymous, nonsense, and frameshift final database. Each SNP's rsID, population frequency (pop. freq.), and DNA variant according to the NCBI's dbSNP, and BLOSUM 62 scores for nonsynonymous variations.

The algorithm can be adapted for other diseases such as various allergies, cardiovascular diseases, or dementia. Ultimately, the algorithm can be generalized and released so that the scientific community can utilize it for other genetic diseases. The anticipated usage for the algorithm is to assist researchers in finding causal variations using NCBI or similar data. Once the causal variations are found, genetic treatment targeting the variations can be produced. There have been similar tools to this algorithm, such as the Broad Institute's MuTect, that helps to identify impactful genetic variations. However, these similar tools do not analyze SNPs and have not been applied to AD, making this tool the first of its kind.

The R script algorithm developed as part of this project helped identify and order the SNPs for *FLG*. Pinpointing the most impactful SNPs is the first step to meaningfully address FLG-related Atopic Dermatitis through precise gene-editing technologies like CRISPR. Our study provides a valuable tool for assisting researchers in establishing potential causal links between SNPs and diseases by prioritizing SNPs that may impact the protein structure.

MATERIALS AND METHODS

The R programming language (version 4.0.3) was used in the Rstudio IDE (Integrated Development Environment). The software packages required were obtained from Bioconductor, an open-source project for biostatisticians to share their software. 'VariantAnnotation' was the primary package used in the project, along with 'GenomicFeatures,' 'GenomicRanges,' 'IRanges,' and 'Biostrings.' For graphing, base R and 'ggplot2' were employed. The starting SNP database was extracted from NCBI's Variation Viewer. The reference genomes came from the UCSC genome browser in the form of 2 R packages: 'TxDb.Hsapiens.UCSC.hg38.knownGene' and 'BSgenome.Hsapiens.UCSC.hg38'. Finally, to analyze the protein for its domains and other features, UniProt and NCBI's Conserved Domain Database was used.

The SNP data was extracted from Variation Viewer Live RefSNPs, dbSNP b154 v2 (range: Chr1: 152,302,165 - 152,325,239), and stored in the Granges object, a genomic data storage class in R. For each SNP (rows), the data stored has three basic categories (columns) that provide the SNP id (rsid), position, and strand information. In addition to the basic categories, additional information, such as variable codon, reference nucleotide, etc. is stored in further metadata columns. The transcript annotations and reference sequence came from the UCSC browser's hg38 assembly. Using the VariantAnnotation package, the R script filtered for SNPs in the coding domain of the protein and added AA information. Then, the code removed the synonymous SNPs since they cause equivalent AA substitution and do not change protein structure. The data at this stage were sent for further processing by the type of SNP (**Figure 6**).

In the first process, the data were sorted into raw frameshift, raw nonsense, and raw nonsynonymous databases. Both nonsense and frameshift SNPs were ordered by location. The code scored and ordered the nonsynonymous SNPs based on the BLOSUM62 substitution matrix (**Figure 7**). Since positive scores indicate similar AA substitution and not a significant change in the protein structure, only non-positive scores were selected for the final analysis (**Figure 4)**. Finally, using NCBI's Conserved Domain Database and the reference sequence, the Ca^{2+} binding region was located.

ACKNOWLEDGMENTS

The authors would like to acknowledge Lake Travis High School for facilitating the research. Furthermore, Dr. Rama Akella and Dr. Prashant Joshi should also be acknowledged for their contributions in reviewing the research for publication and presentations.

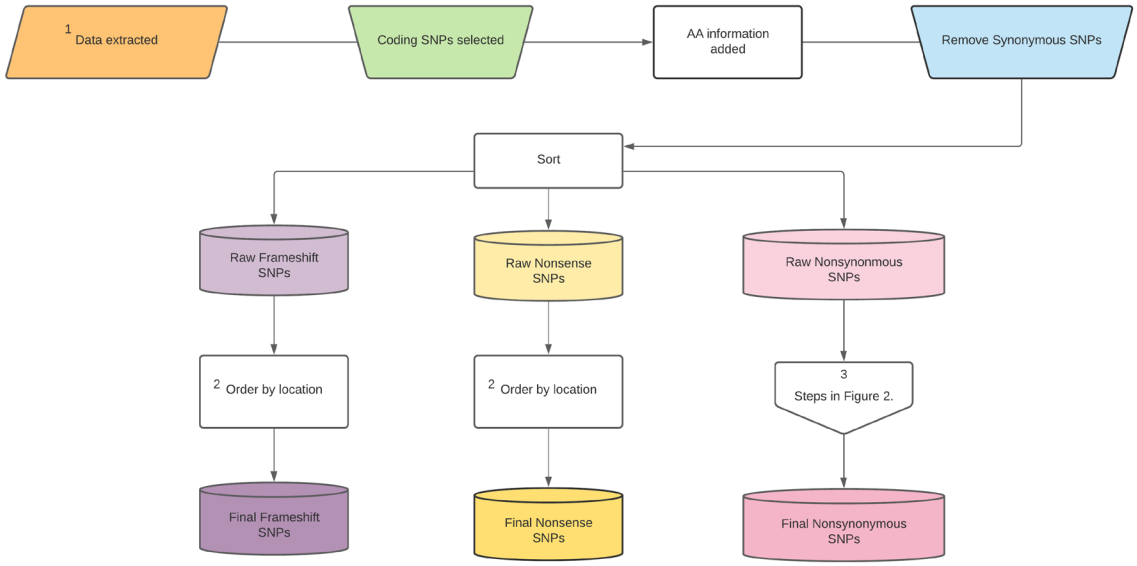

Figure 6: Overall Algorithm workflow. Identification, classification, and ordering of the SNPs in the FLG gene. 1) SNPs were extracted from Variation Viewer Live RefSNPs, dbSNP b154 v2 (range: Chr1: 152,302,165 - 152,325,239). The transcript annotations and reference sequence came from UCSC browser's hg38 assembly. 2) The nonsense and frameshift SNPs were ordered by genomic location 3) Nonsynonymous SNPs were filtered based on their BLOSUM 62 scores.

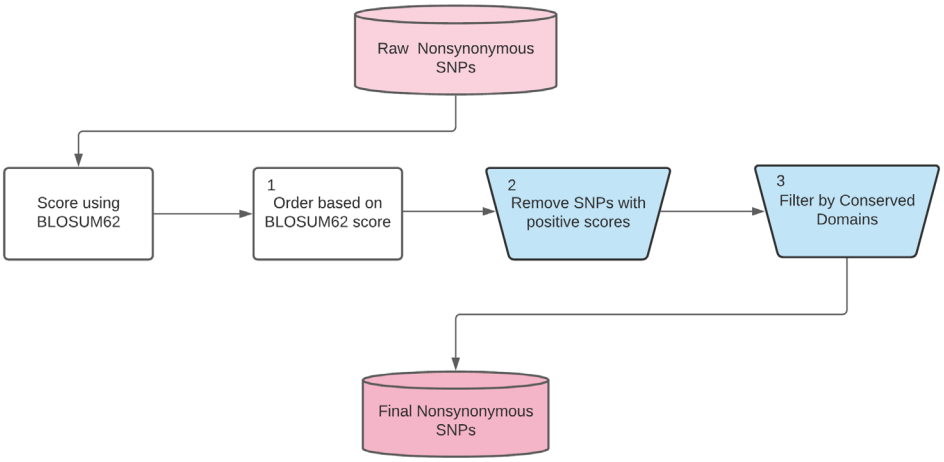

Figure 7: Algorithm workflow for Nonsynonymous SNPs. Scoring of the nonsynonymous SNPs using the BLOSUM62 substitution matrix for filtering and ordering the nonsynonymous SNPs. 1) The lower BLOSUM62 scores indicate possible changes in the protein structure based on the BLOSUM62 substitution matrix since the lower scores represent non-similar AA substitution. Therefore, the SNPs were ordered from the least to the highest BLOSUM62 scores. 2) Since positive scores indicate similar AA substitution and not a significant change in the protein structure, only non-positive scores were selected for the final analysis. 3) Using NCBI's Conserved Domain Database and the reference sequence, the Ca^{2+} binding region was located. The R script identified 25 SNPs that may impact the calcium ion binding region of the native protein because they could impact the production of the Filaggrin protein.

Received: August 8, 2021
Accepted: December 23, 2021
Published: October 12, 2022

REFERENCES

1. Avena-Woods, Carmela. "Overview of atopic dermatitis." *The American journal of managed care* vol. 23,8 Suppl (2017): S115-S123.
2. Skaaby, Tea *et al.* "Associations of filaggrin gene loss-of-function variants and human papillomavirus-related cancer and pre-cancer in Danish adults." *PloS one* vol. 9,6 e99437. 6 Jun. 2014, doi:10.1371/journal.pone.0099437
3. Bickers, David R *et al.* "The burden of skin diseases: 2004 a joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology." *Journal of the American Academy of Dermatology* vol. 55,3 (2006): 490-500. doi:10.1016/j.jaad.2006.05.048
4. Palmer, Colin N A *et al.* "Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis." *Nature genetics* vol. 38,4 (2006): 441-6. doi:10.1038/ng1767
5. Ruff, Samine *et al.* "Prevalence of Cancer in Adult Patients with Atopic Dermatitis: A Nationwide Study." *Acta dermato-venereologica* vol. 97,9 (2017): 1127-1129. doi:10.2340/00015555-2703
6. Brown, Sara J. "What Have We Learned from GWAS for Atopic Dermatitis?." *The Journal of investigative dermatology* vol. 141,1 (2021): 19-22. doi:10.1016/j.jid.2020.05.100
7. Esparza-Gordillo, Jorge *et al.* "Maternal filaggrin mutations increase the risk of atopic dermatitis in children: an effect independent of mutation inheritance." *PLoS genetics* vol. 11,3 e1005076. 10 Mar. 2015, doi:10.1371/journal.pgen.1005076
8. Callaway, E. New concerns raised over value of genome-wide disease studies. *Nature* 546, 463 (2017). doi.org/10.1038/nature.2017.22152
9. Karki, Roshan *et al.* "Defining "mutation" and "polymorphism" in the era of personal genomics." *BMC medical genomics* vol. 8 37. 15 Jul. 2015, doi:10.1186/s12920-015-0115-z
10. Coetzee, Simon G *et al.* "FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs." *Nucleic acids research* vol. 40,18 (2012): e139. doi:10.1093/nar/gks542
11. Sandilands, Aileen *et al.* "Filaggrin's fuller figure: a glimpse into the genetic architecture of atopic dermatitis." *The Journal of investigative dermatology* vol. 127,6 (2007): 1282-4. doi:10.1038/sj.jid.5700876
12. Mucha, Sören *et al.* "Protein-coding variants contribute to the risk of atopic dermatitis and skin-specific gene expression." *The Journal of allergy and clinical immunology* vol. 145,4 (2020): 1208-1218. doi:10.1016/j.jaci.2019.10.030
13. Dakal, Tikam Chand *et al.* "Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in IL8 gene." *Scientific reports* vol. 7,1 6525. 26 Jul. 2017, doi:10.1038/s41598-017-06575-4
14. Murata, Susumu *et al.* "Interleukin-8 Levels in the Stratum Corneum as a Biomarker for Monitoring Therapeutic Effect in Atopic Dermatitis Patients." *International archives of allergy and immunology* vol. 182,7 (2021): 592-606. doi:10.1159/000512965
15. Ortiz, Romina A, and Kathleen C Barnes. "Genetics of allergic diseases." *Immunology and allergy clinics of North America* vol. 35,1 (2015): 19-44. doi:10.1016/j.iac.2014.09.014