IOSUD - Universitatea Politehnica Timişoara
Şcoala Doctorală de Studii Inginereşti

IN SILICO SOLUTIONS FOR RESEARCHING GENOMIC VARIANTS AND PATTERNS APPLYING SYSTEMS ENGINEERING METHODS

PhD Thesis – Abstract

To obtain a PhD degree at
Politehnica University Timişoara
In Systems Engineering

Author inf. Cristian-Grigore ZIMBRU

Thesis supervisor Prof.univ.dr.ing. Ioan SILEA
Month 6 year 2020

In the thesis entitled „In silico solutions for researching genomic variants and patterns applying systems engineering methods” are presented a series of methods for processing genetic information obtained after secondary analysis of data generated by DNA sequencing equipment[1].

The thesis is structured in three parts with a total of six chapters (Fig. 1) distributed as follows:

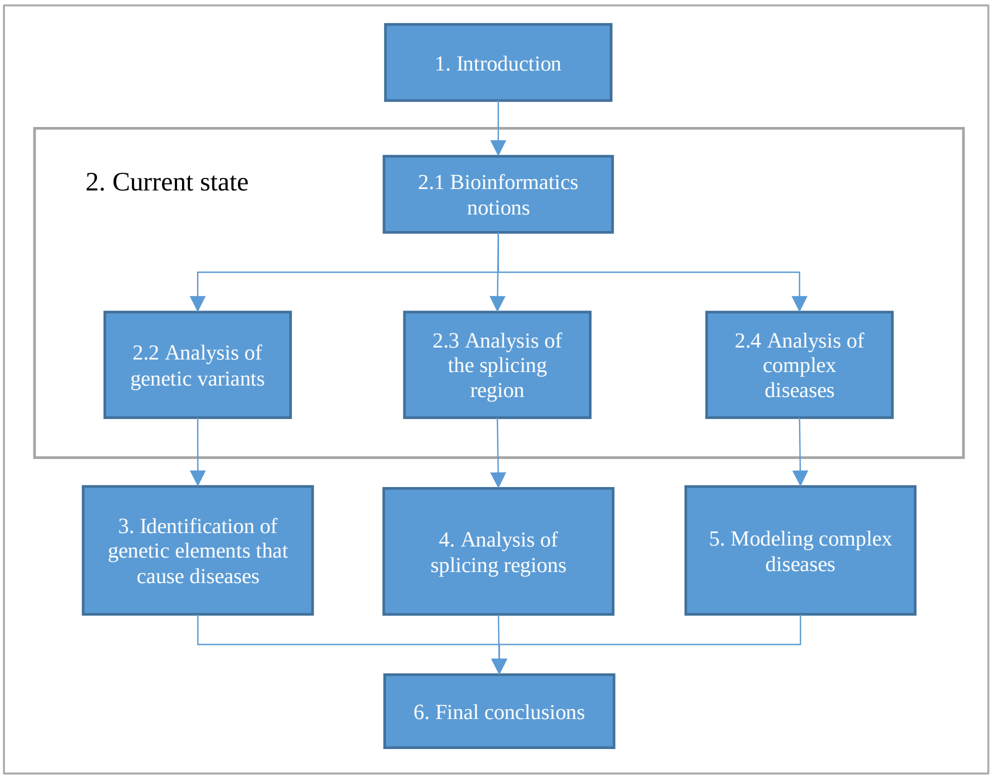
1. Introduction
2. Current state
2.1 Bioinformatics notions
2.2 Analysis of genetic variants
2.3 Analysis of the splicing region
2.4 Analysis of complex diseases
3. Identification of genetic elements that cause diseases
4. Analysis of splicing regions
5. Modeling complex diseases
6. Final conclusions

Fig. **Error! No text of specified style in document.** Thesis structure

1) the first part contains chapters 1 and 2 which are intended to present the topic and issues addressed in the thesis;

[1] HiSeq 2500, MiSeq, Sanger 3730xl.

2) the main part consists of chapters 3, 4 and 5 which present the solutions proposed for solving the problems dealt with in each chapter;
3) in the last part the final conclusions and personal contributions are presented.

The concrete aim of the thesis is to develop and use computer methods (*in silico*) that are able to identify the genetic structures that cause diseases. The contributions aim to highlight the existence of genetic variants and patterns for: (*i*) monogenic diseases, (*ii*) diseases caused by splicing, but also (*iii*) prediction of complex diseases (steatosis case study).

We will continue to refer, very briefly, the essential aspects that are found in each chapter of the thesis.

The **first chapter** presents the field of the thesis, the opportunity of writing the thesis and the pursued purpose and objectives. The thematic orientation of the thesis aims in principle:

- Determining the disease using information about phenotype and genotype;
- Identification of the disease or the genetic variant causing it based on the genetic signature;
- Reducing the number of genetic variants using certain filtering strategies;
- Determination of the variation of the splicing signal following genetic changes;
- Identification of splicing regions;
- Modeling of complex diseases based on genotype.

For the development of the research and the realization of the work, major objectives with related sub-objectives were proposed, presented below.

1) Development of a method or workflow capable of indicating a small number of genetic variants that explain the characteristics of a patient's phenotype, which includes:
   a) Identification of applications and research that address the issue of disease detection based on phenotype or genotype;
   b) Reducing the number of possible conditions of a patient depending on the targeted gene panel and its genotype;
   c) Identification of the optimal strategy for the use of in silico predictors for the detection of pathogenic genetic variants;
   d) Extraction of tolerance intervals for genetic variants, depending on the target gene panel.
2) Identification of splicing regions using computational models of the splicing sequences and of the enhancer and silencer signals of this process. This involves performing the following tasks:
   a) Review of applications that identify splicing regions and intensity of splicing signals;
   b) Analysis of the performance of the methods for detecting the splicing regions and proposing models for their identification and extraction from the intronic regions;
   c) Development of an algorithm that allows the calculation of the splicing signal intensity;
3) Generating a computational model for the prediction of a complex condition (steatosis) using a set of genetic markers, is the third objective and requires the following:
   a) Identification of genetic markers relevant to steatosis.
   b) Studying the performance of models generated by the methods used in machine learning

for the prediction of complex diseases;

c) Proposing models for predicting steatosis using methods based on a set of statistical models or methods.

In practical terms, the results of the research should facilitate the identification of the condition, in the case of undiagnosed patients, or the identification of the causative genetic variants, in the case of patients whose condition is known but the genetic reason is not known. Specifically, for the first objective, the series of algorithms (methods) will be applied on the genotype of a person, more precisely on the exonic variants, and their result will consist in the list of possible diseases. The second objective is to consider the genetic changes that affect the splicing process. The methods developed in this section aims to identify the changes that occurred in the splicing process. The ultimate goal is to generate models that should be able to predict with more than 80% accuracy the genetic predisposition to steatosis or even its presence.

Chapter 2 is structured in four subchapters. The first subchapter (2.1) aims to present some introductory notions to familiarize the reader (engineer, computer scientist) with elements of molecular biology and bioinformatics. In the case of molecular biology, some terms are explained that will be used quite frequently in the content of the thesis, such as; uninucleotide polymorphism, locus, gene, reference genome etc. After the presentation of the genetic terms, we move on to the presentation of the concepts of bioinformatics. The technology used for DNA sequencing is described, and then the secondary analysis workflow is presented. The last part of the subchapter presents some information in the field of machine learning and how to evaluate the performance of a model generated by such methods.

The second subchapter (2.2) deals with the issue of identifying genes and genetic variants that cause diseases. The types of diseases are cataloged according to [1] and the relativity of the information is presented when it comes to cataloging a genetic variant in databases such as ClinVar [2]. Silico prediction tools are used to identify pathogenic variants, which have associated pathogenicity scores for each genetic variant. It starts with the presentation of silico predictors such as: SIFT [3], CADD [4], PROVEAN [5], etc. It also presents a number of international initiatives concerned with the processing and cataloging of genetic information, such as the UK 100k Genomes project [6]. Below are the methods by which the filtering and selection of genetic variants is done. A first filter, which can be applied to these variants, is the elimination of those who have a high frequency in the population, ideally those in the population where the patient comes from. For the frequency in the population of genetic variants, one can use databases such as 1000 Genomes or Gnomad [7]. Also, in this subchapter are presented applications proposed in the literature for the analysis of genetic variants, such as Phevor2 [8], eXtasy [9], etc.

Subchapter (2.3) contains information on the pre-mRNA splicing. At the beginning of the subchapter is a brief presentation of the splicing process and how are the results analyzed in genetics. Although, initially the variants that were not part of the exons of a gene were ignored, in recent years specialists have begun to give more importance to variants that affect the splicing process and to study their role in the manifestation of diseases. To improve the results in terms of splicing regions, specialists use prediction software tools. The splicing sequence contains three important components: (1) the branch point (BRS), the pyrimidine tract and the acceptor site. Of these, the BRS and the pyrimidine tract are not well preserved [10]-[12], only the acceptor site is preserved, being formed by the nucelotides AG. For the prediction of the splicing region, a series of applications are presented such as SplicePort, Automated Splice Site Analyzes, MaxEnt, etc. In addition, the regulatory elements (cis-acting) of the splicing are presented. These elements are represented by: the sequences between intron and exon, intronic enhahacers and silencers and exonic enhahacers and silencers.

The last subchapter (2.4) presents a short study on steatosis and genetic markers that indicate the presence of this condition. Hepatic steatosis is the accumulation of fat in the liver tissue and two forms of it are known (one due to alcohol consumption and another due to other factors). The literature indicates a number of genetic factors that may predispose a person to this condition. The main candidate for this condition is the rs738409 polymorphism, which will be discussed in Chapter 5. The paper [13] presents an analysis of genetic findings that are associated with fatty liver disease, including rs738409. Also, in this subchapter are presented methods in the field of machine learning used in the field of complex diseases. Initially, the studies that treated the problem of complex diseases were association studies. An alternative to this method is the Bayesian methods or the methods used in machine learning. Decision trees, ensemble methods, neural networks can easily identify the patterns that appear in multidimensional data sets.

The topics in **Chapter 3** focus on a number of methods that help improve the detection of genes and genetic variants that cause disease. In most cases, geneticists have at their disposal several types of information about patients, such as the file with the genetic variants, the patient's symptoms and possibly the family history. Based on this information, doctors need to identify which genes are responsible for the ailments the patient suffers from. The first method presented involves an association of the characteristics of the phenotype with the elements of the genotype, thus resulting in a list of possible diseases. Several databases available online have been used to develop an automated system for identifying disease-causing genes. The Human Phenotype Ontology (HPO) database was used to obtain the list of symptoms, and the databases used to obtain the list of diseases were Online Mendelian Inheritance in Man (OMIM), Orphanet and DECIPHER. For the hierarchy of genes, respectively of variants, the use of similarity coefficients are proposed. After calculating the weights and similarity coefficients, the patient's genotype is applied together with the characteristics of the phenotype to determine the list of possible diseases. This method has satisfactory but not complete results.

In order to improve the performance of the previously presented method, *in silico* predictors can be used. Although in the literature there are various performance tests of these predictors, unfortunately they are not performed on the same database. Therefore, a performance test with a common database was performed. The database used was ClinVar, and the metrics used were accuracy, F_1 score and average between specificity and sensitivity. Regarding the results, CADD and DANN identified the most pathogenic variants and had associated scores for more than 95% of the data set. The disadvantage of these predictors was that they had a relatively low specificity. REVEL, MetaSVM and PolyPhen-2 HVAR had the best overall performance, calculated with the arithmetic mean between specificity and sensitivity. As a routine for the correct classification of SNPs, pathogenic genetic variants could be determined with high-sensitivity instruments (CADD and DANN) and then balanced predictors can be used (REVEL, MetaSVM, PolyPhen) to prioritize them. [14].

Subchapter 3.4 dealt with the detection of quality and quantity errors of the genetic variants identified following the sequencing process. A method for identifying errors using tolerance intervals is applied. The intervals presented are indicative, but at the same time they can be used as a reference for error detection. It is recommended that each laboratory use such methods to determine the quality of sequencing. In addition to possible errors, these intervals may signal certain causes of the condition such as inbreeding.

In **Chapter 4** we analyzed the elements of the RNA sequence that are part of the splicing process, namely the sequence and the splicing signals. These elements were processed from the perspective of DNA.

The study carried out in subchapter 4.1 identified some splicing regions that have two or more sequences that correspond to branch point (BRS) of the spliceosome. The experiment was structured in two stages. The first stage consisted in defining a model for the splicing sequences, based on works from the literature [10] as well as on the results of the analysis of over 11000 splicing sequences from chromosome 21. The second step was to use the definition of the model generated in step 1 to identify pseudo-splicing regions in the intronic area. The results obtained in the first stage indicated a degree of redundancy of the BRS regions for certain exons, which may be due to the pyrimidine region or which may have a biological significance - this requires a more detailed investigation. The branch point regions were often located near positions 16 and 28, upstream of the three prim exon. The model of a splicing region, following the results obtained in the first stage, consists of: (1) a BRS region which has the YTnAy model, (2) an AG acceptor region and (3) a pyrimidine region which consists of 75 % pyrimidine nucleotide bases with a length of at least 17 bases [15]. Using this model, intronic sequences were also identified *in silico* that are similar in structure to the splicing regions. The biological role of these sequences cannot be validated *in silico,* but can be tested *in vitro* or *in vivo* experiments [16].

The study in section 4.2 consisted of analyzing the splicing sequences in the Homo Sapiens Splice Site Dataset database using various methods. Following the analysis, a series of statistical information on the structure of splicing sequences were presented and a series of models were generated that were meant to validate these regions. The models initially presented were based on equations generated from the structure of the splicing regions (order of nucleotides, tuples, etc.). The accuracy of the prediction of these models was between 70% and 80%. Although it is a decent level, the targeted accuracy was around 90%, the performance of the MaxEnt method. In the last part of the chapter, a method was proposed for the detection of splicing regions based on the distance from the neighboring sequences. A number of methods were analyzed to calculate the distance, and the one finally chosen was Needleman-Wunsch. Using this method, a computational analysis was performed to determine an optimal sequence length and an optimal number of neighbors. The results indicated that the sequence should be 20 nucleotides long and the ideal number of neighbors is nine. Using these values it was possible to obtain an accuracy of 85.61%.

The purpose of the study in subchapter 4.3 was to develop a method to annotate VCF files with information about splicing signal variations. In the first phase, the databases with nucleotide sequences which were considered as signals for the splicing process were gathered.

These sequences, in their initial form, had various criteria for calculating the signal strength, which limited their simultaneous use. Therefore, a series of equations were developed to determine the signal strength of a sequence. Next, to calculate the amplitude difference between two sequences, the initial one and the one containing the genetic modification, the average intensity of the non-zero position vector related to the sequence was calculated. The amplitude calculation is performed both for the splicing enahancer signal and for the splicing silencer signal. The general direction is given by the analysis of these two components. The validation of the method was performed on a database that contains the genetic sequences (normal and modified) and that contains the indication of the behavior of the splicing process. In addition, the results correspond to the detailed information indicated by the Human Splicing Finder. The method can be used to filter and prioritize genetic variants [17].

A series of models for the prediction of steatosis based on a list of genetic markers are presented in **Chapter 5**. The first subchapter presents the used materials and the list of genetic variants used to make the models. In the second subchapter, a descriptive analysis of the records in the database is performed according to the targeted genetic variants. Also in this subchapter is analyzed the correlations between genetic variants and the degree of steatosis.

The third subchapter presents a series of prediction models for different degrees of this condition. The first model investigated is the one generated using the Stohastic-Gradient Descent (SGD) method. For the individual stage prediction, the SGD model has an average accuracy of about 70%, but for the prediction of the five states of steatosis, simultaneously, this value decreases significantly, being 25%. Next, decision trees were used for multi-class prediction. To determine the optimal configuration, a number of parameters were analyzed. The average accuracy score for multiple-stage steatosis was 30%. Also in this subchapter, an attempt was made to model the steatosis stage using a set of decision trees (Random Forest). The accuracy in this case was 36%, marginally higher than in the case of a single decision tree.

In subchapter four, the degree of complexity is reduced by eliminating the stages of steatosis and replacing them with the simple presence of the disease. In this situation, using the decision trees, it was possible to obtain a model that managed an accuracy of 91%, the average score being 81% [18]. The same performance was obtained in the case of the set of decision trees. As presented in the literature [19], [20], both methods generated patterns indicating that the SNP rs738409 is associated with steatosis.

Decision trees performed better in the second phase, when the steatosis stage was reduced. This may indicate that the sample size used was too small to model the output, but was optimal for determining whether the pathology is present. In addition, the difference between male and female subjects was not taken into account. Some SNPs may be more or less relevant based on gender. One configuration, with the best accuracy for decision trees, was the use of the Gini Index function with MDL prunnig and random sampling.

In subchapter five, a method was developed that generates prediction models according to the frequency of occurrence and "expertise" of each SNP. Using this method, the condition of steatosis being binary, an accuracy of 82% was obtained. This model has a lower performance than that of decision trees and overall prediction models. However, the method has several advantages such as the generation of voting maps that make it much easier to identify the relationships between SNPs and the condition. In fact, a number of relationships have been highlighted for steatosis. For example, SNPs rs2167444 and rs7848 on the heterozygous SCD gene appear to have an affinity for the zero stage of steatosis; or the SNPs on the ABCB4 gene appear to indicate an affinity for steatosis state 2.

Chapter 6 presents the conclusions, personal contributions and future directions of development. The main personal contributions:

1. Development of a method for the determination of pathogenic genetic variants according to the characteristics of the phenotype and the variants detected in patients;
2. Carrying out a study to identify the best method of using *in silico* predictors in filtering genetic variants. Presentation of results and suggesting prioritization strategies;
3. Propose a method for determining the tolerance ranges used in the detection of sequencing errors. This method can also be used to quickly identify the causes of diseases such as inbreeding;
4. Carrying out a study of all introns on chromosome 21 to generate a statistical model of the splicing sequence;
5. Implementation of a computer method for the detection of parasitic splicing sequences in intronic regions;
6. Development of a method for calculating the variation of the splicing signal in case of DNA sequence modification;
7. Identification of redundant sequences for branch point in the splicing region;
8. Presentation of a method for the detection of splicing regions according to the distance between the target sequence and the neighboring sequences using the Needleman-Wunsch algorithm;;
9. Carrying out a statistical study that presents the structure of the splicing region;
10. Determination of in silico models using decision trees for the prediction of the presence of steatosis, respectively determination of its stage based on genotype;
11. Development of a method for predicting the presence of steatosis and its stage based on the frequency of genetic variants;
12. Implementation and validation of the aforementioned methods using data from the Center for Genomic Medicine (UMFT).

At the same time, in the thesis are referred, out of the **14** papers published by the author of the thesis as first author or co-author, the **5** that validated the research results. In short, according to the level of impact at which the various scientific results were communicated, the grouping of published works is as follows:

- 2 papers in indexed journals Web of Science (ISI) – **IF 7.2**
- 3 works in volumes of scientific manifestations (proceedings) indexed Web of Science (ISI)
- 2 papers in BDI indexed journals (IEEE Xplore) - submitted for integration in Web of Science (ISI)
- 7 papers in the volumes of scientific events.

The consistency of the research results from the paper "Splice Site Pattern Analysis and Identification of Similar Sequences in the Deep Intron Areas of Human Chromosome 21" led to the 3rd prize at the 2017 EHB conference, and following the paper "Detection of high-risk intron areas that can cause splicing errors” was awarded a *Young Scientist Fellowship*.

The thesis has 146 pages, of which: 116 pages structured in 6 chapters, 10 pages of bibliography and 20 pages dedicated to the annexes. The paper contains 64 figures and 151 bibliographic titles. Some of the contributions presented have been published in scientific papers in which the author of the thesis is the author or co-author, and others will be the subject of future papers and collaborations.

[1] Jonathan Pevsner, *Bioinformatics And Functional Genomics, 3rd edition*. 2015.

[2] M. J. Landrum *et al.*, “ClinVar: public archive of interpretations of clinically relevant variants,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862-868, Jan. 2016, doi: 10.1093/nar/gkv1222.

[3] P. C. Ng and S. Henikoff, “SIFT: predicting amino acid changes that affect protein function,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[4] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, “CADD: predicting the deleteriousness of variants throughout the human genome,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D886–D894, Jan. 2019, doi: 10.1093/nar/gky1016.

[5] Y. Choi and A. P. Chan, “PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels,” *Bioinformatics*, vol. 31, no. 16, pp. 2745–2747, Aug. 2015, doi: 10.1093/bioinformatics/btv195.

[6] M. Caulfield *et al.*, “The National Genomics Research and Healthcare Knowledgebase.” Aug. 21, 2019, doi: 10.6084/m9.figshare.4530893.v5.

[7] K. J. Karczewski *et al.*, “Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes,” *bioRxiv*, p. 531210, Aug. 2019, doi: 10.1101/531210.

[8] M. V. Singleton *et al.*, “Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families,” *Am. J. Hum. Genet.*, vol. 94, no. 4, pp. 599–610, Apr. 2014, doi: 10.1016/j.ajhg.2014.03.010.

[9] A. Sifrim *et al.*, “eXtasy: variant prioritization by genomic data fusion,” *Nat. Methods*, vol. 10, no. 11, pp. 1083–1084, Nov. 2013, doi: 10.1038/nmeth.2656.

[10] K. Gao, A. Masuda, T. Matsuura, and K. Ohno, “Human branch point consensus sequence is yUnAy,” *Nucleic Acids Res.*, vol. 36, no. 7, pp. 2257–2267, Apr. 2008, doi: 10.1093/nar/gkn073.

[11] D. A. Bitton *et al.*, “LaSSO, a strategy for genome-wide mapping of intronic lariats and branch-points using RNA-seq,” *Genome Res.*, p. gr.166819.113, Apr. 2014, doi: 10.1101/gr.166819.113.

[12] A. J. Taggart, A. M. DeSimone, J. S. Shih, M. E. Filloux, and W. G. Fairbrother, “Large-scale mapping of branchpoints in human pre-mRNA transcripts *in vivo*,” *Nat. Struct. Mol. Biol.*, vol. 19, no. 7, pp. 719–721, Jul. 2012, doi: 10.1038/nsmb.2327.

[13] S. Sookoian and C. J. Pirola, “Genetics of Nonalcoholic Fatty Liver Disease: From Pathogenesis to Therapeutics,” *Semin. Liver Dis.*, vol. 39, no. 2, pp. 124–140, May 2019, doi: 10.1055/s-0039-1679920.

[14] C. G. Zimbru, N. Andreescu, A. Albu, A. Chirita-Emandi, A. Stanciu, and M. Puiu, “Performance Evaluation of in Silico Predictors for the Classification of ClinVar Variants,” in *2019 E-Health and Bioengineering Conference (EHB)*, Nov. 2019, pp. 1–4,

doi: 10.1109/EHB47216.2019.8969963.

[15] C. G. Zimbru *et al.*, “Splice site pattern analysis and identification of similar sequences in the deep intron areas of human chromosome 21,” in *2017 E-Health and Bioengineering Conference (EHB)*, Jun. 2017, pp. 145–148, doi: 10.1109/EHB.2017.7995382.

[16] Cristian Zimbru, Nicoleta Andreescu, Adela Chirita-Emandi, Antonius Stanciu, Ioan Silea, Maria Puiu, “Detection of high-risk intron areas that can cause splicing errors,” *Adv. Lect. Course Syst. Biol.*, p. p 74, Mar. 2016.

[17] C. G. Zimbru, A. Albu, N. Andreescu, A. Chirita-Emandi, and M. Puiu, “Determining Splicing Signal Variation in Humans by Analyzing the Regulatory Splicing Motifs,” in *2019 E-Health and Bioengineering Conference (EHB)*, Nov. 2019, pp. 1–4, doi: 10.1109/EHB47216.2019.8969983.

[18] C. G. Zimbru, N. Andreescu, A. Chirita-Emandi, I. Silea, M. Puiu, and M. D. Niculescu, “Analysis of decision tree performance in predicting the relationship between a scored outcome and multiple single nucleotide polymorphisms,” in *2017 E-Health and Bioengineering Conference (EHB)*, Jun. 2017, pp. 57–60, doi: 10.1109/EHB.2017.7995360.

[19] P. J. Meffert *et al.*, “The PNPLA3 SNP rs738409:G allele is associated with increased liver disease-associated mortality but reduced overall mortality in a population-based cohort,” *J. Hepatol.*, vol. 68, no. 4, pp. 858–860, Apr. 2018, doi: 10.1016/j.jhep.2017.11.038.

[20] S. Grimaudo *et al.*, “PNPLA3 rs738409 C>G variant predicts occurrence of liver-related events and death in non-alcoholic fatty liver,” *Dig. Liver Dis.*, vol. 51, pp. e5–e6, Feb. 2019, doi: 10.1016/j.dld.2018.11.045.