Ben Gurion University of the Negev

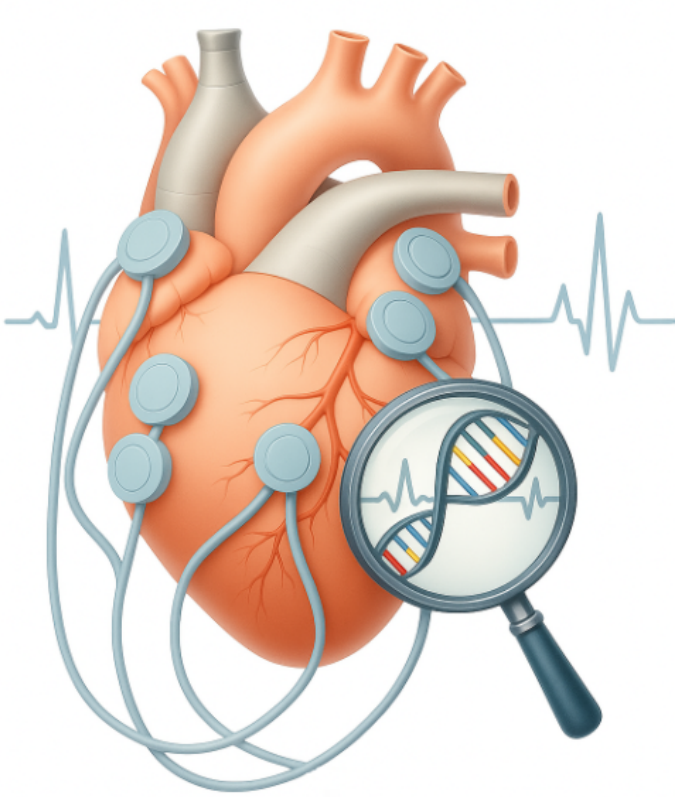
Faculty of Engineering Science

Department of Software and Information Systems

Deciphering The Genomic Landscape Through ECG Signal Analysis: During Rest

זיהוי מאפיינים גנטיים בסיגנל פיזיולוגי במנוחה

June 2025



Submitted By:

Stav Farajun
Selin Neomi Ivshin
Orel Jarahian

Supervisors:

Dr. Nadav Rappoport
Mr. Natan Lubman

1. Abstract

At the heart of this project lies an ambitious goal: to uncover how the human genome shapes cardiac function through the analysis of ECG signals using advanced machine learning technologies. Rather than relying on predefined clinical metrics, we developed an innovative model that extracts latent features from physiological signals and integrates them with genetic data.
The study was based on the UK Biobank database, which contains ECG recordings alongside genetic information for approximately a thousand participants. A thorough preprocessing phase was conducted, including the removal of noisy and invalid files using the Wavelet DB4 method, resulting in a clean, high-quality dataset. Subsequently, a 1D Residual CNN Autoencoder model was trained to generate compact and informative representations of the signals.
These representations were then used to perform a genome-wide association study (GWAS), identifying genetic SNPs associated with cardiac function. The findings were further analyzed using bioinformatic tools such as HaploReg, to investigate coding and regulatory regions, and STRING, to examine gene networks and uncover relevant biological interactions.
As part of this project, we built an innovative scientific and technological framework that can serve as a foundation for developing personalized prediction tools, enabling early diagnosis of heart conditions, and advancing our understanding of genome-physiology relationships. This marks a significant step toward a future of personalized medicine, where real-time biological signals are integrated with each individual's unique genetic profile, at the intersection of life sciences and artificial intelligence.

בלב הפרויקט ניצבת מטרה שאפתנית: לחשוף כיצד הגנום האנושי מעצב את פעילות הלב, באמצעות ניתוח של אותות א.ק.ג בטכנולוגיות של למידת מכונה. במקום להסתמך על מדדים רפואיים מוגדרים מראש, פותח מודל חדשני המחלץ מאפיינים נסתרים מתוך אותות פיזיולוגיים ומצליבה אותם עם מידע גנטי.
העבודה התבססה על מאגר ה־UK Biobank, הכולל הקלטות ECG לצד מידע גנטי עבור כאלף נבדקים. תהליך הניקוי הקפדני כלל סינון רעשים וקבצים בעייתיים באמצעות Wavelet DB4, ליצירת בסיס נתונים איכותי ונקי. בהמשך, אומן מודל 1D Residual CNN Autoencoder ליצירת ייצוגים קומפקטיים ומשמעותיים מהסיגנלים.
מאפיינים אלו שימשו לביצוע מחקר גנומי (GWAS), שבחן אילו SNPs גנטיים מקושרים לפעילות הלב. ממצאים אלו נותחו באמצעות כלים ביואינפורמטיים כמו HaploReg להערכת אזורי קידוד ורגולציה, ו־STRING לניתוח רשתות גנים ולהבנת אינטראקציות ביולוגיות רלוונטיות.
במסגרת הפרויקט יצרנו תשתית מדעית וטכנולוגית חדשנית, המסוגלת להוות בסיס לפיתוח כלי חיזוי מותאמים אישית, לקידום אבחון מוקדם של מחלות לב, ולהעמקת ההבנה של קשרי גנטיקה־פיזיולוגיה. זהו צעד משמעותי לעבר עתיד של רפואה מותאמת אישית, המשלבת בין נתונים ביולוגיים בזמן אמת לבין הפרופיל הגנטי הייחודי של כל אדם - בדיוק במקום שבו מדעי החיים פוגשים את הבינה המלאכותית.

Contents

1. Abstract **1**

2. Introduction **3**

3. Literature Review **4**

3.1 Background Review 4

3.2 Related Work 5

4. Research Purposes **12**

5. Research Methodology **12**

5.1 ECG Signal Preprocessing 12

5.1.1 Data Extraction 12

5.1.2 Baseline Wander Removal 13

5.1.3 Denoising via Discrete Wavelet Transform (DWT) 13

5.1.4 Quality Control 14

5.2 1D Residual CNN Autoencoder Architecture 14

5.2.1 Model Input 14

5.2.2 Encoder Architecture 14

5.2.3 Latent Representation 15

5.2.4 Decoder Architecture 15

5.2.5 Baseline Wander Removal 15

5.3 Model Training 16

5.3.1 Signal Normalization 16

5.3.2 Training Protocol 16

5.4 Genome-Wide Association Study (GWAS) 17

5.4.1 Genotype–Phenotype Association Across ECG Morphology Features 17

5.4.2 Covariates for Confounder Control 17

5.4.3 Association Testing Procedure 17

5.5 Finding Significant SNPs and Creating Manhattan Plots 18

5.6 Gene Mapping and Interaction Analysis 18

6. Model Development and Experimental Evaluation **19**

6.1 Experiments on FFT - Filtered Signals 19

6.2 Experiments on DWT - Filtered Signals 20

7. Results **22**

8. Conclusions **28**

9. Future Directions and Challenges **27**

10. Bibliography **29**

11. Appendices **30**

2. Introduction

The relationship between human genetics and cardiac function has long been of interest to researchers aiming to understand the biological mechanisms behind heart activity. Traditionally, studies in this area have relied on predefined clinical measurements or expert-annotated features, limiting the scope of discovery to known markers. However, the growing availability of large-scale physiological and genomic datasets, such as those provided by the UK Biobank, opens the door to new data-driven approaches that can uncover previously hidden patterns. This project is motivated by the need to move beyond manual signal interpretation and to develop automated, scalable tools that can bridge the gap between raw physiological signals and genomic information. While previous studies have explored the genetic basis of heart disease using diagnostic criteria or aggregate metrics (e.g., heart rate, QT interval), few have attempted to directly link raw ECG signals to genetic variation using modern machine learning. To address this gap, we propose a novel pipeline that integrates advanced signal processing, deep representation learning, and genome-wide association analysis. The process began with ECG signals being cleaned and denoised, using Wavelet DB4 to remove artifacts and ensure data quality. Then, a 1D Residual CNN Autoencoder was trained to learn compact and informative latent representations from the raw signals. These learned features were used in a GWAS to identify genetic variants (SNPs) correlated with signal-derived phenotypes. Our analysis revealed significant associations between the extracted features and genes involved in cardiac ion channels, particularly potassium and calcium transport - key players in electrical conduction in the heart. These findings highlight the potential of combining physiological and genetic data to uncover meaningful biological insights.

This document's structure includes the following: Introduction, Literature Review, Research purposes, Research methodology, Experiments, Results, Bioinformatics Analysis, Conclusions, and Future Work.

3. Literature Review

3.1 Background Review

Autoencoders are a type of neural network that can learn to represent data in a simpler, low-dimensional way without needing labeled data. They have two main parts: the encoder, which takes the input and compresses it into a smaller, more meaningful representation (called the latent space), and the decoder, which tries to recreate the original input from this compressed version. The model learns by reducing the difference between the original input and the reconstructed output, ensuring that the most important features of the data are captured.

CNN Autoencoders are a specific type of autoencoder that use convolutional layers. These are especially useful for working with structured data like images, signals, or time series data. The encoder in a CNNAE uses convolutional and pooling layers to identify and extract important patterns, while the decoder uses methods like transposed convolutions or upsampling to rebuild the original data. In addition, CNNAE are particularly valuable in signal analysis, where high-quality data is crucial for accurate analysis and decision-making.

In the context of our project, we will examine the value of CNNAE in the processing and analysis of electrocardiograms (ECGs).

Electrocardiograms (ECGs) are non-invasive, graphical evaluations of car-diovascular electrical activity. The technique involves using external electrodes strategically placed at specific locations on patients' chest, arms, and legs, providing clinicians with visual representations of heart electrical signals and rhythms.

ECG is a widely available test in cardiovascular evaluations, and models can be trained to identify and detect hidden patterns of disease from those electrical signals.

Recently, different machine learning approaches, including support vector machines and convolutional neural networks, have been applied to identification of heart rhythm disorder and differentiation of genotypes using 12-lead ECG. [1] [2]

The dataset we will use in this project includes resting-state ECG measurements only. To the best of our knowledge, no prior work has been conducted in the field of ECG and its genetic associations specifically comparing activity under exercise to rest.

By studying this link, we hope to understand how genetic factors affect the heart's electrical activity, enhance cardiovascular health, and, through simple tests, detect potential heart-related diseases. To explore this connection further, it is essential to consider the genetic under-pinnings that drive variations in ECG parameters.

Single Nucleotide Poly-morphisms (SNPs) and Genome-Wide Association Studies (GWAS) are fundamental concepts in modern genetics research, particularly in understanding the genetic basis of complex traits like cardiovascular diseases and ECG parameters.

SNPs are the most common type of genetic variation among humans, occurring approximately once every 1,000 nucleotides. These variations represent differences in a single DNA building block at a specific position in the genome. For a variation to be considered a SNP, it must occur in at least 1% of the population. Scientists have identified over 600 million SNPs in populations worldwide, making them invaluable tools for genetic research.

In the context of cardiovascular research, SNPs have been linked to various ECG parameters, including QT interval, PR interval, QRS duration, and heart rate. For instance, SNPs in genes encoding cardiacion channels and nitric oxide synthase 1 adaptor protein (NOS1AP) have been associated with ECG QT-interval duration. [3]

GWAS is a powerful approach to identifying genetic variations associated with specific traits or diseases. GWAS examine SNPs across the entire genome in large groups of individuals to find genetic variations associated with a particular trait. This method has revolutionized the field of genetics by allowing researchers to identify novel genetic loci associated with complex traits without prior hypotheses about which genes might be involved.

The connection between SNPs, GWAS and ECG traits is profound and has significantly advanced our understanding of cardiac electrical function. GWAS have been instrumental in identifying genetic variants associated with various ECG parameters, such as PR interval, QRS duration, QT interval, and heart rate. These studies involve analyzing SNPs across the entire genome in large cohorts to find associations between genetic variations and specific ECG traits. For instance, a meta-analysis of GWAS results from approximately 30,000 samples identified over 130 significant loci associated with these ECG parameters, including 17 loci for PR interval, 13 for QRS duration, 15 for QT interval, and 7 for heart rate. [4]

3.2 Related Work

Large-scale biobanks, like the UK Biobank and Biobank Japan, offer extensive datasets that combine genomics, biomedical information, and health records from hundreds of thousands of participants, creating opportunities to explore complex traits and disorders. [5] [6] [7]

Despite this, utilizing this high-dimensional data effectively for genetic discovery remains a significant challenge. There are various approaches to working with high-dimensional clinical data, each with its own specific limitations. For example, using **principal component analysis (PCA)** assumes linear relationships within the data, which can potentially reduce accuracy. Additionally, employing supervised machine learning models for phenotype label prediction may cause the model to focus solely on signals associated with the specific target trait.

Previous studies have proposed deep learning frameworks that use low-dimensional embeddings to uncover genetic associations, leveraging tools to quantify clinically relevant features not yet standardized or automated.

For example, Taedong Yun et al. introduce **REGLE (REpresentation learning for Genetic discovery on Low-dimensional Embeddings)**, a frame-work that applies deep learning techniques, specifically Variational Autoen-coders (VAEs), to generate low-dimensional, interpretable embeddings from high-dimensional clinical data (HDCD). [8]

REGLE was tested on data from **spirograms and photoplethysmograms (PPGs), representing respiratory and circulatory systems**, respectively. It demonstrated significant advantages over conventional methods by discovering novel genetic loci and improving polygenic risk scores (PRS) for diseases such as asthma, chronic obstructive pulmonary disease (COPD), hypertension, and systolic blood pressure.

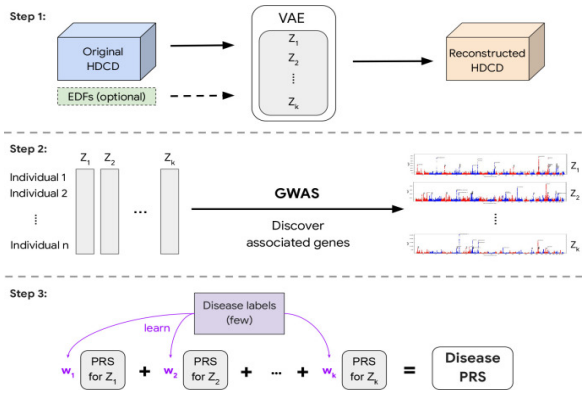


Figure 1: Overview of REGLE. The process begins with learning a low-dimensional embedding using a VAE, optionally conditioned on EDFs. Next, GWAS is performed on the learned coordinates and EDFs. Finally, a linear model combines latent coordinate PRSs to generate the disease-specific PRS.

The methodology involved three primary steps: learning embeddings using convolutional VAEs, performing GWAS on the embedding coordinates, and constructing disease-specific PRSs by linearly combining the PRS of individual embedding coordinates. This architecture also allowed for the optional incorporation of expert-defined features (EDFs), **enabling the model to learn residual signals beyond these predefined features.**

In terms of results, REGLE identified a substantial number of novel genome-wide significant loci. For example, 11% of the loci associated with SPINCs (spirogram embeddings) were previously unreported for lung function, while 56% of loci associated with PLENCs (PPG embeddings) were novel for cardiovascular traits. These discoveries highlight REGLE's potential to uncover previously hidden genetic insights. Furthermore, the PRSs derived from RE-GLE significantly outperformed those based on traditional EDFs in predicting disease risks.

While Yun et al. applied their REGLE framework to spirograms and PPGs, demonstrating the efficacy of VAEs in reducing complex data into low-dimensional, herita-

ble embeddings, **our project applies this approach to ECG signals recorded at rest.** Unlike the use of VAEs in REGLE, our project employs a **CNNAE and integrates signal filtering as a preprocessing step to en-sure that the input data for the autoencoder is optimized for feature extraction.**

This focus on dynamic data may provide novel insights into the genetic underpinnings of cardiovascular function and health, complementing and expanding the findings of REGLE by demonstrating the broader applicability of deep learning techniques across diverse physiological signals.

Building on the innovative use of deep learning for genetic discovery as demonstrated in the work by Yun et al. (2023), the study by Sieliwonczyk et al. (2024) provides a complementary perspective by **employing VAEs to identify genetic determinants of electrocardiographic features.** [7]

This research explores the potential of unsupervised learning to uncover novel insights into cardiac electrophysiology, utilizing ECG data to extract latent factors (LFs) that represent comprehensive and interpretable features. The authors applied their model to over one million ECG records from secondary care, with external validation using UK Biobank data, to successfully identify novel genetic loci associated with ECG morphology.

The study utilized a VAE with one-dimensional convolutional layers to extract meaningful LFs from ECG data. These layers effectively captured spatial de-pendencies in the signals, while a **beta-parameter** optimized the trade-off between **reconstruction accuracy and feature disentanglement**, ensuring the LFs were both interpretable and biologically relevant. Median-beat ECGs were processed using the BRAVEHEART software, an open-source tool that standardized signal extraction, reduced noise, and ensured scalability across datasets, including over one million ECGs for training and validation.

The VAE demonstrated superior reconstruction accuracy (Pearson correlation of 0.94-0.95) compared to earlier models. GWAS of the latent factors identified 120 significant single-nucleotide polymorphisms (SNPs) across 65 unique loci, of which 42% were novel. Rare-variant analyses further revealed associations with seven genes, including NEK6 and IL17RA, highlighting the capacity of VAEs to capture both common and rare genetic variations in cardiac traits. Additionally, phenotypic analyses demonstrated the superiority of the LFs in detecting correlations with cardiac disease markers and echocardiographic features compared to conventional ECG metrics.

Despite its strengths, the study has limitations that align with gaps identified in previous works. The dataset predominantly represented European ancestry, limiting its applicability to diverse populations.

Based on the methodology and findings presented by Sieliwonczyk et al, **our study extends this work using resting-state ECG analysis.** While their study focused on resting ECGs, our project introduces a critical dimension by analyzing dynamic ECG data captured during cycling sessions. As an alternative to the VAE discussed here, we propose a CNNAE with a different architecture.

Expanding the exploration of genetic influences on cardiac electrophysiology, the next study **focuses on cardiovascular genomics and computational biology**, examining genetic influences on ECG phenotypes [9].

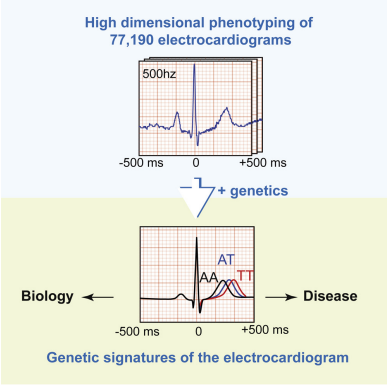


Figure 2: The Genetic Makeup of the Electrocardiogram - Graphical Abstract

It addresses the problem of a limited understanding of how genetic variants influence the entire ECG cycle. Traditional genome-wide association studies (GWAS) focus on isolated ECG segments, such as the PR interval or QT duration, which limits the ability to gain insights into the genetic determinants of the complete cardiac conduction process. To overcome this, **the study uses high-dimensional phenotyping of 77,190 ECGs from the UK Biobank, segmenting each ECG into 500 spatial-temporal data points to represent the entire cardiac cycle**. This approach, combined with GWAS, iden-tifies genetic loci (a specific, fixed position on a chromosome where a particular gene or genetic marker is located) linked to ECG morphology.

The study introduces a new method of analyzing ECG as a high-dimensional phenotype rather than relying solely on predefined traits. It identifies genetic signatures across the cardiac cycle, revealing unique associations between ge-netic loci and specific ECG characteristics. The integration of clustering and causal inference techniques enhances the understanding of the genetic architec-ture underlying cardiac electrophysiology. This approach captures the complex-ity of ECG data, offering detailed insights into genetic influences throughout the cardiac cycle. The study identifies over 300 genetic loci, including several novel ones not previously associated with ECG traits, highlighting its significance in advancing cardiovascular genomics.

However, the study's **reliance on a predominantly European ancestry dataset limits its generalizability to other populations**, as genetic variations and their effects can differ across ethnic groups. Additionally, it does not include functional validation of the identified loci and excludes rare genetic variants, focusing instead on common variants identified through GWAS. **Key findings from the research highlight over 300 genetic loci associated with ECG morphology, including loci linked to conditions.**

New loci were identified that are associated with specific ECG features, providing deeper insights into the mechanisms underlying cardiac dis-eases. **Clus-**

tering analyses reveal distinct genetic effects on various ECG segments, such as the Q wave and T wave, which are linked to im-portant biological processes like ion-channel activity. The research also shows that ECG features corresponding to regions of high electrical activity are highly heritable.

This research demonstrates the potential of combining GWAS with high-dimensional phenotyping to uncover genetic influences on physiological signals. Its methods align with our project goals, which explore connections between ECG signals and genetic variations. The use of spatial-temporal data and advanced computational techniques provides a framework adaptable for deep learning approaches.

The gap between the article and the project lies in the methodologies and aims of each. The article focuses on analyzing high-dimensional ECG data. It emphasizes spatiotemporal analysis, clustering, and more but does not incorporate advanced computational modeling.

In contrast, our project incorporates latent space modeling of ECG data to extract nuanced features, filtering significant SNPs, and mapping gene expression to physiological states for deeper biological understanding. This allows for a deeper exploration of the relationships between genetic variations and phenotypic traits, moving beyond the statistical associations in the article to a framework that connects genomic data with biological functions.

While the focus of the previous study lies in exploring genetic influences on ECG phenotypes, **another intriguing line of research examines the inte-gration of genetic markers with phenotypic traits in a different con-text—facial recognition.**[10]

Although it is not directly relevant to the connection between ECG and genetics, it provides insights into integrating genetic markers and phenotypic traits. This research uniquely applied genetic data to explore facial recognition using face-to-DNA classifiers.

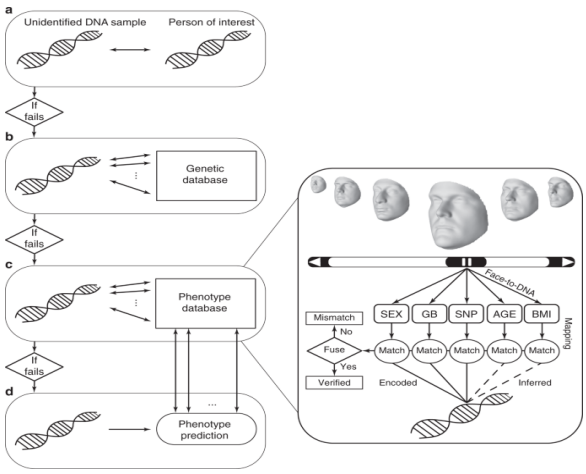


Figure 3: The proposed method links DNA to phenotypes for identification. If direct matching fails, face-to-DNA classifiers compare features like sex and age to a phenotype database. As a last resort, predicted traits may rely on public recognition, though DNA phenotyping is still limited.

This investigation introduces a novel approach to biometric authentication by linking genetic information to facial identity. **Instead of predicting facial features from DNA, it uses face-to-DNA classifiers to infer genetic traits like sex, genomic background and individual genetic loci from 3D facial shape data.** The method is applied to diverse and homogeneous cohorts, improving specificity and recognition rates through GWAS and multiple classifiers.

Results revealed strong recognition accuracy in diverse populations, with **true matching rates of 83% and 80% in verification mode**. However, it faces challenges in identifying individuals within more homogeneous groups. Combining multiple molecular features enhances performance, whereas individual predictors like sex alone show limited effectiveness. The methodology avoids direct facial reconstructions from DNA and leverages 3D facial shape analysis, though it is constrained by dataset structure, predefined genetic markers, and privacy concerns.

This innovative approach is particularly relevant to research aiming to bridge phenotypic and genetic data, providing insights into integrating genetic markers and phenotypic traits through advanced classification techniques. While the study focuses on facial recognition, the methodological framework shares similarities with our project, exploring the connection between physiological signals, such as ECG, and genetic variations. Both approaches seek to understand how genetic factors influence observable traits, using computational methods to uncover latent relationships. **However, our project differs in its focus on physiological monitoring data and deep learning models to extract patterns from genetic and physiological data.** Unlike the article's re-liance on structured GWAS and classifiers, the project utilizes deep learning to uncover nonlinear and complex interactions, broadening the scope beyond biometric applications.

Following the investigation into the integration of genetic markers with phenotypic traits in facial recognition, **another significant area of research shifts focus to the genetic factors influencing cardiac function.** [11]

The research explores how genetics and genomics explain the QT interval in ECGs, which measures heart electrical activity. It focuses on identifying genetic factors affecting heart function to address gaps in understanding why some individuals are more prone to conditions like ventricular arrhythmia and sudden cardiac death.

Using genome-wide association studies (GWAS) on a dataset of over 250,000 individuals, the study identifies genetic variants linked to heart function. It also uses rare-variant analysis and bioinformatics tools like polygenic risk scores (PRS) and pathway enrichment to uncover over 200 genetic regions associated with QT, JT, and QRS intervals. The multi-ancestry dataset and large scale of the study make it impactful, revealing pathways like insulin receptor signaling and energy metabolism that could guide new treatments for heart rhythm disorders.

The study's strengths include its robust dataset and discovery of novel genetic pathways and rare variants, expanding knowledge of heart function. Advanced bioinformatics links genetic factors to physical traits and drug targets, open-ing possibilities for therapeutic interventions. However, it does not explore how these genetic fac-

tors directly influence heart function, focusing instead on statis-tical associations. Additionally, underrepresentation of some populations limits the findings' generalizability.

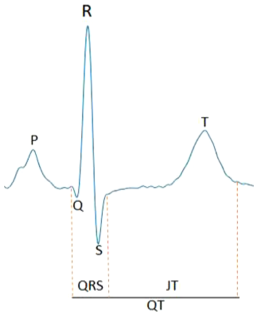


Figure 4: Annotation of an example ECG signal. QRS duration and the JT interval approximate the time periods for ventricular depolarization and repolarization on the surface ECG. The entire segment from onset of the Q wave to end of the T wave is the QT interval

Key findings of the study include the identification of 176 genetic regions related to the QT interval, 156 to the JT interval, and 121 to QRS duration, including a unique male-specific gene on the X-chromosome. By analyzing data from over 250,000 individuals, the study revealed that QT and JT intervals are linked to pathways like insulin signaling and energy metabolism, while QRS duration is associated with connective tissue and cell growth processes. Rare genetic variants were connected to heart disease genes, such as KCNQ1 and KCNH2, demonstrating how both rare and common genetic changes influence heart function. Additionally, polygenic risk scores (PRS) linked these genetic findings to conditions like atrial fibrillation and sudden cardiac death.

The article is relevant to our project as it provides insights into the genetic architecture of ECG traits, such as QT, JT, and QRS intervals and their associations with genetic variations. This aligns with the project's goal of connecting physiological signals with genetic variations. Additionally, the findings on polygenic risk scores and their connections to cardiac conditions offer valuable benchmarks and data points that our project can build upon to enhance its ability to uncover non-linear and complex relationships.

The gap between the article and the project lies in methodology and scope. The article uses GWAS and statistical methods to identify associations between genetic variations and ECG traits focusing on linear relationships and pathway enrichment from static data. **In contrast, the project integrates deep learning to uncover non-linear, complex relationships, utilizing latent space modeling to extract detailed and biologically relevant features from ECG data.** Additionally, while the article focuses on existing datasets, the project includes gene expression mapping to differentiate physiological states, enabling deeper insights and potential real-time monitoring of genetic influences on health.

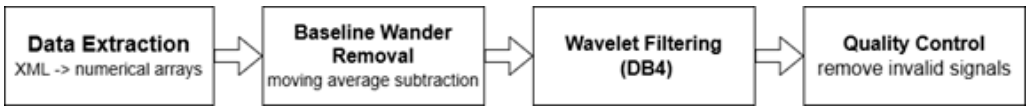
4. Research Purposes

Our research aims to bridge the gap between complex physiological signals, such as ECG recordings, and genetic variations by applying advanced machine learning techniques. The developed pipeline compresses and interprets raw signal data, extracting latent representations that may capture hidden genomic influences on phenotypic traits. This approach enables the discovery of novel genetic associations that traditional, manually engineered features may overlook - contributing to a deeper and more data-driven understanding of human biology. In the long term, this approach offers the potential to go beyond traditional diagnostics and support clinical decision-making in a more individualized way. By identifying subtle genomic patterns in physiological signals, this project lays the groundwork for precision medicine - where treatments and risk assessments are customized based on a person's genetic profile.

5. Research Methodology

5.1 ECG Signal Preprocessing

Before analyzing the resting-state ECG signals, we first applied a step-by-step cleaning process to improve the quality of the data. This preprocessing aimed to remove common issues like noise and slow signal shifts (called baseline drift), while keeping the important features of the heart signal intact. The main steps in this process are described below:



5.1.1 Data Extraction

ECG signals were stored in `.xml` files, with each file corresponding to a specific subject and lead (there are 12 leads per individual). To organize the data for analysis, we first parsed the filenames to extract a unique (subject ID, lead) key for each recording. During this process, we encountered duplicate recordings for certain (ID, lead) combinations. These typically included two versions: one ending in `2_0.xml` and another in `3_0.xml`. Upon inspection, we found that files ending in `3_0.xml` were relatively rare and corresponded to recordings from a different year than those labeled `2_0.xml`. Since `2_0.xml` files were more consistent and representative across the dataset, we chose to retain them and exclude the `3_0.xml` duplicates to ensure temporal consistency and maximize data uniformity.

5.1.2 Baseline Wander Removal

Baseline wander, a low-frequency artifact in ECG recordings, is caused by factors like breathing, movement, and electrode impedance. Prior to denoising, baseline drift was eliminated using a moving average subtraction technique. A reflective-mode uniform moving average filter was applied to each signal with a window of 0.5 seconds (corresponding to 250 samples at a 500 Hz sampling rate). The filtered trend was then subtracted from the original signal to remove low-frequency baseline fluctuations.

5.1.3 Denoising via Discrete Wavelet Transform (DWT)

In this stage, a denoising procedure based on the Discrete Wavelet Transform was applied to the raw ECG signals. This method enables the decomposition of a signal into multiple frequency bands while maintaining temporal localization, making it particularly effective for preserving relevant features and eliminating noise.
The core principle of wavelet-based denoising is that many real-world signals, including biological waveforms like ECGs, have sparse representations in the wavelet domain. Most of the signal's essential structure is concentrated in a few large-magnitude wavelet coefficients, while noise is typically spread across many low-magnitude coefficients. By applying a thresholding operation (commonly referred to as wavelet shrinkage) low-magnitude coefficients associated with noise are removed and the denoised signal is reconstructed using the inverse wavelet transform. [12]

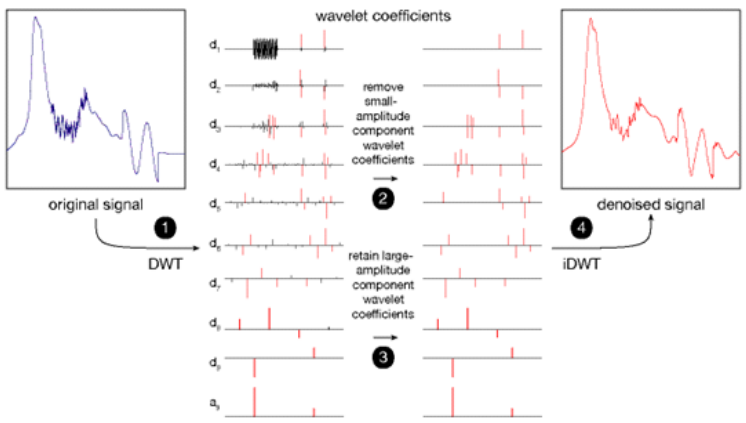


Figure 5: Illustration of Wavelet-Based Denoising Process
The original signal is decomposed via Discrete Wavelet Transform (DWT), followed by thresholding of small-magnitude coefficients. The denoised signal is reconstructed using the inverse DWT (iDWT).

In this project, we used Daubechies-4 (Db4) wavelets, which are well suited for biomedical signals due to their ability to capture sharp transients such as QRS complexes. A four-level wavelet decomposition was performed. The threshold was calculated based on the widely adopted VisuShrink method, which assumes additive white Gaussian noise. This choice offers a strong balance between noise suppression

and signal fidelity, as supported by prior studies in biomedical signal processing literature.

5.1.4 Quality Control

After preprocessing, each signal was validated for numerical integrity. Signals containing 'NaN' or 'Inf' values were flagged as invalid and excluded from further analysis. Valid signals were saved in NumPy binary format (.npy) for efficient downstream processing.

5.2 1D Residual CNN Autoencoder Architecture

This architecture was selected following extensive experimentation with various network configurations, including multiple encoder-decoder structures, convolutional depths, and latent space dimensions. Through iterative tuning of hyperparameters and architectural components, the final model was chosen based on its ability to meet two key objectives:

- **Minimizing the dimensionality** of the latent space to enable efficient downstream genetic analysis
- **Achieving high-quality signal reconstruction**, ensuring that the autoencoder preserves essential morphological features of the original ECG waveform.

5.2.1 Model Input

The model receives a one-dimensional ECG segment of length 5000, corresponding to 10 seconds of signal at a 500 Hz sampling rate. The input is representing a single-lead trace.

5.2.2 Encoder Architecture

The encoder comprises three sequential residual convolutional blocks, each including:

- Two 1D convolutional layers with kernel size 7 and L2 regularization
- Batch normalization layers to stabilize training
- LeakyReLU activation functions for non-linearity
- Dropout (rate = 0.2) applied after each block for regularization

Each encoder block down-samples the input using a stride of 2 in the convolutional layer, which progressively reduces the temporal resolution while increasing the number of feature channels (**32 → 64 → 128**). The residual connections help preserve important signal features and improve gradient flow during training, particularly in deeper architectures.

5.2.3 Latent Representation

Following the residual blocks, the feature maps are flattened and passed through a fully connected layer, yielding a latent vector of dimension 32. This vector captures the compressed essence of the ECG morphology and serves as the input for downstream genomic analysis.

5.2.4 Decoder Architecture

The decoder reconstructs the original ECG signal from the compressed latent representation by progressively upsampling and refining the signal:

- It begins with a fully connected (dense) layer, which reshapes the latent vector back into the dimensions of the final convolutional feature map produced by the encoder.
- A dense layer that reshapes the latent vector into the original convolutional feature map size
- LeakyReLU activation functions for non-linearity
- Dropout (rate = 0.2) applied after each block for regularization
- The decoder then passes the signal through three upsampling blocks, which progressively increase the temporal resolution while **reducing the number of feature channels** (**128 → 64 → 32**). Each block consists of:
  - An **UpSampling1D** operation to increase temporal resolution
  - A **1D convolutional layer** with a kernel size of 7 to refine the signal
  - **Batch normalization** to stabilize learning
  - A **LeakyReLU activation function** to introduce non-linearity
- Finally, the output is passed through a **1D convolutional layer with linear activation**, which reconstructs the ECG trace to its original input length and scale.

5.2.5 Baseline Wander Removal

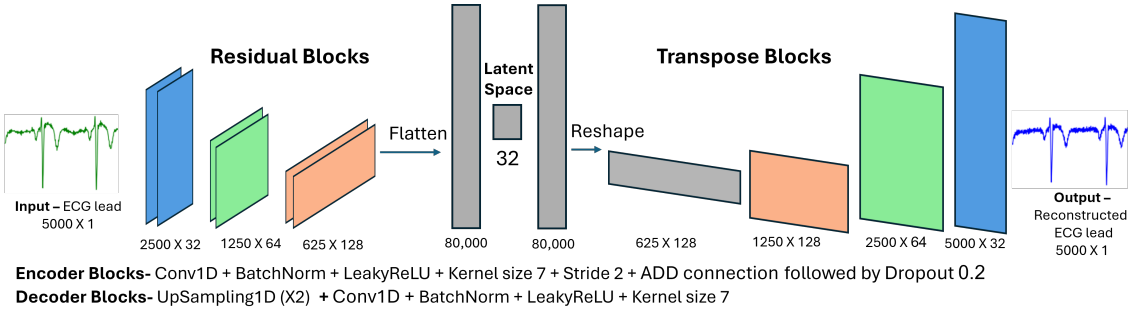
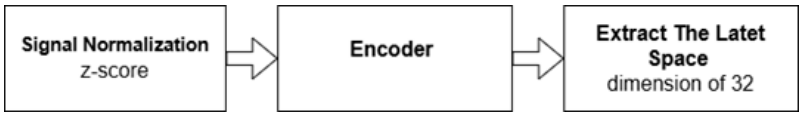


Figure 6: Architecture of the 1D Residual Convolutional Autoencoder for ECG Signal

Illustration of the encoding and decoding process applied to a single ECG lead, from raw input to reconstructed output, highlighting dimensional transformations at each stage:



5.3 Model Training

Following the design of the 1D Residual Convolutional Autoencoder, the model was trained to reconstruct ECG signals using a standardized training and validation protocol.

5.3.1 Signal Normalization

To ensure consistent signal dynamics across the dataset, all signals were normalized using z-score standardization, calculated from the training set:

$$X_{\text{norm}} = \frac{X - \mu}{\sigma}$$

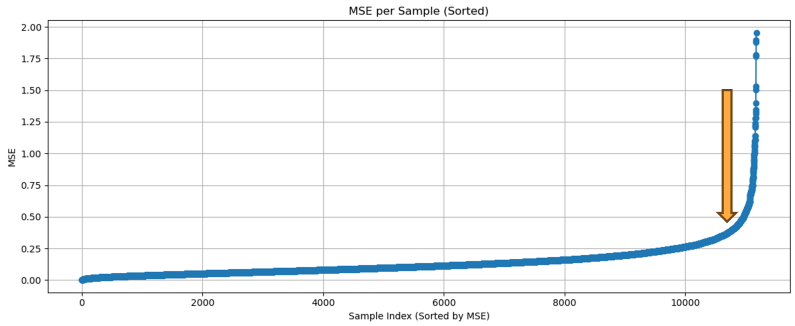
Where μ is the global mean and σ is the global standard deviation of the training signals. This normalization process was applied identically to both training and validation data using the training statistics.

5.3.2 Training Protocol

The model was compiled with the **Mean Squared Error** loss function and optimized using the **Adam optimizer** (learning rate $= 10^{-3}$).
To improve generalization and prevent overfitting, two callbacks were used during training:

- **EarlyStopping:** Monitored validation loss with a patience of 10 epochs, restoring the best weights upon termination.
- **ReduceLROnPlateau:** Dynamically reduced the learning rate by a factor of 0.5 if the validation loss plateaued for 5 consecutive epochs.

Training was performed over **a maximum of 100 epochs**, using a **batch size of 32**. Validation data was used to monitor performance and adjust learning dynamics in real-time. After running the autoencoder model, latent space files with an MSE greater than 0.4 were excluded (approximately 3% of the signals) from the rest of the pipeline to remove outliers.



5.4 Genome-Wide Association Study (GWAS)

To identify genetic variants associated with ECG-derived features, we used **PLINK 2.0**, a standard tool in genetic epidemiology for conducting GWAS. PLINK supports efficient regression analysis across millions of SNPs and allows for the inclusion of covariates to control for confounding factors.
In this section, we used a SLURM job array to efficiently execute multiple GWAS tests.

5.4.1 Genotype–Phenotype Association Across ECG Morphology Features

The input to the GWAS consisted of 32 latent features derived from compressed ECG representations, with each feature representing a distinct aspect of ECG morphology. Association testing was conducted separately for each of the 12 ECG lead, represented by a latent space of 32 features. The analysis utilized binary PLINK files (.bed, .bim, .fam) from the UK Biobank genotype dataset. Each autosomal chromosome (chromosomes 1–22) was tested independently against every latent feature and SNP.

5.4.2 Covariates for Confounder Control

To account for demographic, population structure, and technical confounders, a comprehensive set of covariates was incorporated into the regression model. These included:

- **Age** at assessment
- **Genetic sex**
- **Population structure** via the first 10 genetic principal components (PCs)
- **Assessment center identifiers**, representing the location where samples were collected. These were included as dummy-coded variables for UK Biobank centers 11001 through 11021.

This adjustment ensures that observed associations are not confounded by ancestry, age, sex, or recruitment site.

5.4.3 Association Testing Procedure

Each run of PLINK was executed with the following configuration:

- Model: Linear regression
- Phenotypes: All 32 latent features included
- Missing phenotype values been Excluded
- **PLINK output:** Separate results files were generated per lead, per chromosome and per feature

5.5 Finding Significant SNPs and Creating Manhattan Plots

This stage of the pipeline was responsible for identifying statistically significant genetic associations between SNPs and each ECG latent feature, and for visualizing these results using Manhattan plots.
The data was loaded and preprocessed:

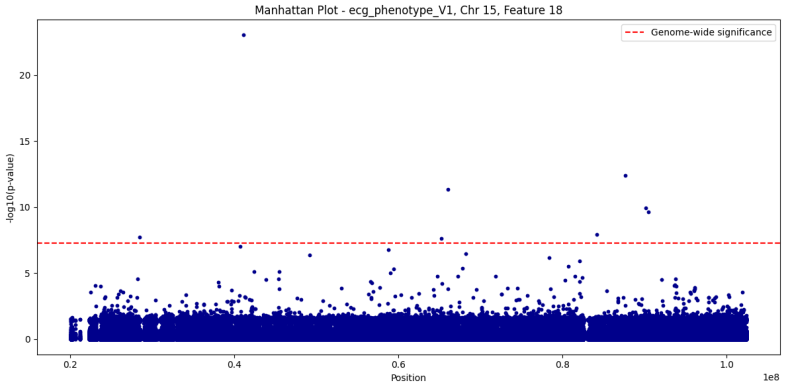
- P-values with missing entries (NA) were filtered out.
- A new column **-log10(P)** was calculated to facilitate significance thresholding and plotting.

To determine statistical significance, we applied a genome-wide significance threshold of $\mathbf{-\log_{10}(p) > 8.8}$, corresponding to a p-value of 5×10^{-8}. This threshold was derived by adjusting the conventional genome-wide threshold of 5×10^{-8} to account for the 32 independent latent features extracted from the ECG data. Specifically, we used a **Bonferroni-like correction**:

$$\frac{5 \times 10^{-8}}{32} = 1.56 \times 10^{-9}, \quad -\log_{10}(1.56 \times 10^{-9}) \approx 8.8$$

This conservative correction controls for the increased risk of false positives due to multiple hypothesis testing across the 32 ECG features.

If any significant associations were detected, a Manhattan plot was generated to provide a visual overview of the genetic signals across the genome. These plots enable quick identification of loci exhibiting strong associations. Subsequently, all significant SNPs identified for each feature-chromosome combination were consolidated into a single CSV file. This file served as the input for the downstream stages of the analysis.



5.6 Gene Mapping and Interaction Analysis

Significant SNPs were mapped to gene names using either HaploReg / g:Profiler, depending on the number of associations obtained. Specifically, HaploReg was used when only a small number of significant results were available, while g:Profiler was preferred for larger SNP sets. In practice, the majority of analyses involved extensive results and were therefore processed using **g:Profiler**. Following gene mapping, we performed protein–protein interaction analysis using the **STRING** database, applying MCL clustering (an algorithm that finds groups of proteins that are naturally related to each other) with an inflation parameter of 2.0. This parameter controls the granularity of clustering, where higher values result in more stringent and tightly defined clusters. The analysis was restricted to the species Homo sapiens (human) to ensure biological relevance.

6. Model Development and Experimental Evaluation

To evaluate and optimize our approach, we conducted a series of experiments spanning both the data preprocessing stage and the model development phase. During preprocessing, we compared two signal denoising techniques - Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT) - to determine which method yields cleaner ECG signals. In the model-building phase, we explored a wide range of architectural and training configurations for the autoencoder, including various normalization methods, activation functions, layer types, kernel sizes, number of epochs, and loss functions. These experiments aimed to identify the most effective design choices for learning meaningful representations of ECG signals.

6.1 Experiments on FFT - Filtered Signals

A summary of the key experiments conducted on ECG signals preprocessed using the FFT denoising technique:

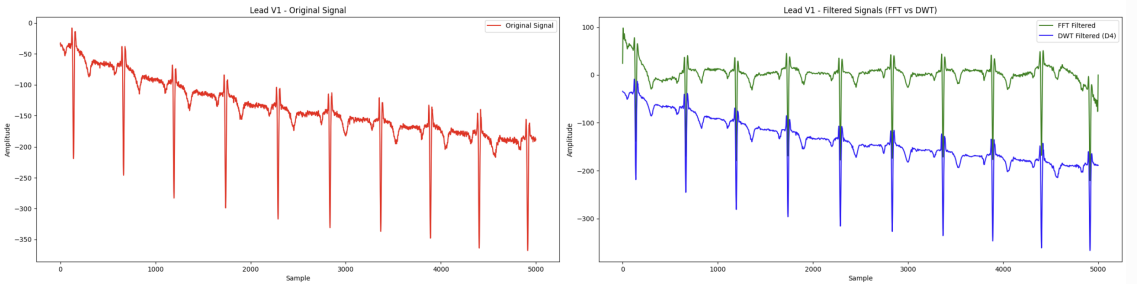
- **Optimizer:** The model was trained using the **Adam optimizer**, chosen for its efficiency and consistent performance in related studies.
- **Loss function: Mean Squared Error** was used, as it is the standard choice for signal reconstruction tasks.
- **Regularization: L2 regularization** with a weight of **0.001** was applied to reduce overfitting.
- **Normalization:** No normalization was applied during this phase, as preliminary tests showed it did not improve performance or GWAS outcomes.
- **Activation functions:** Several activation functions were tested:
  - **Tanh** produced smooth but blurry reconstructions.
  - **Leaky ReLU** yielded sharper and more realistic waveforms, and was therefore preferred.
- **Skip connections:** We experimented with skip connections **between encoder and decoder**. While they improved reconstruction loss, they introduced **data leakage** that reduced latent compression quality.
  As a result, this technique was excluded from the final FFT-based models.
- **Latent space and training epochs:** The autoencoder was trained with varying numbers of epochs and multiple latent space dimensions to identify the optimal configuration for effective signal representation.

Latent Space Dim	Skip Connections	Activation	Layers in encoder/decoder	Kernel Sizes	Filters
200	Without	ReLU	6× [Conv1D + BatchNorm + MaxPooling] + Dense	7,7,5,5,3,3	32,32,16,16,8,4
100	Without	tanh	4× [Conv1D + BatchNorm + MaxPooling] + Dense	7,5,3,3	32,16,8,4
100	Yes (Add) between the Encoder to Decoder	LeakyReLU	5× [Conv1D + BatchNorm + MaxPooling] + Dense	9,7,5,3,2	32,16,8,4,2
48	Yes (Add) between the Encoder to Decoder	LeakyReLU	5× [Conv1D + BatchNorm + MaxPooling] + Dense	9,7,5,3,2	32,16,8,4,2
24	Yes (Add) between the Encoder to Decoder	tanh	4× [Conv1D + BatchNorm + MaxPooling] + Dense	9,7,5,3	32,16,8,4

Table 1: Summary of autoencoder configurations and results on FFT-cleaned ECG signals

6.2 Experiments on DWT - Filtered Signals

To improve the preprocessing stage, an essential step that prepares the input data for effective autoencoder training, we opted to use Discrete Wavelet Transform (DWT) instead of Fast Fourier Transform (FFT). Although FFT-based filtering yielded lower reconstruction error (i.e., lower MSE), visual inspection revealed that DWT-preserved signals maintained richer morphological features of the original ECG waveforms. While DWT denoising resulted in higher reconstruction loss, it may facilitate the extraction of more meaningful latent representations, which are critical for downstream applications such as Genome-Wide Association Studies. Accordingly, we evaluated both filtering methods not only by reconstruction accuracy, but also by their ability to produce informative and biologically relevant latent spaces.



In these experiments, we evaluated the impact of various normalization strategies, regularization techniques, and skip connection designs on the performance of our autoencoder models. We compared three normalization types:

- **Per-signal normalization**, (z-score applied independently per signal)
- **Global normalization**, (z-score using the training set mean and std)
- **No normalization**

While per-signal normalization improved reconstruction accuracy (i.e., lower MSE), it removed global variance between signals and significantly reduced the biological informativeness of the latent representations, resulting in poor GWAS performance. In contrast,

global normalization preserved inter-signal variance and yielded more meaningful GWAS associations.

We further applied:

- **Dropout** (rate = 0.2) and **L2 regularization** ($\lambda = 0.0001$) to reduce overfitting.
- **Skip connections** applied independently in the encoder and decoder. We experimented with different skip connection methods: **element-wise addition** and **concatenation**.

Our results highlight that architectural and preprocessing choices - particularly normalization and skip connection strategy-strongly influence not only reconstruction loss but also the utility of the learned representations for downstream genomic analysis.

Latent Dim	Normalization Type	Skip Connections	Activation	Layers in encoder/decoder	Kernel Sizes	Filters	Epochs	Test Loss
32	Without	Yes (Add) in Encoder & Decoder separately	LeakyReLU	3 residual blocks [Conv1D + BN + LeakyReLU, L2] + Dense	7,7,7	32,64,128	100	1560
32	Per-signal	Yes (Add) in Encoder & Decoder separately	LeakyReLU	3 residual blocks [Conv1D + BN + LeakyReLU, L2] + Dense	7,7,7	32,64,128	100	0.187
64	Per-signal	Yes (Concatenate) in Encoder & Decoder separately	LeakyReLU	5× [Conv1D + BatchNorm + MaxPooling] + Dense	9,7,5,3,3	32,16,8,4,2	45	0.51
32	**Global (Z-score)**	**Yes (Add) in Encoder only**	**LeakyReLU**	**Encoder: 3 residual blocks x[Conv1D + BN + LeakyReLU, L2] + Dropout (0.2) + Dense. Decoder: 3x[Upsampling + Conv1D + BN + LeakyReLU, L2]**	**7,7,7**	**32,64,128**	**100**	**1.28**
32	Without	Yes (Add) in Encoder & Decoder separately	LeakyReLU	3 residual blocks [Conv1D + BN + LeakyReLU, L2 + Dropout (0.2)] + Dense	7,7,7	32,64,128	100	1909

Table 2: Summary of autoencoder configurations and results on DWT-cleaned ECG signals

The selected model, trained on DWT - filtered signals with a latent dimension of 32 and global z-score normalization, was chosen as the final architecture due to its ability to produce the highest number of genome-wide significant associations, despite not achieving the lowest reconstruction loss (test MSE of 1.28).

This model demonstrated the best trade-off between reconstruction quality and biological informativeness, and was therefore used for the final latent feature extraction phase prior to GWAS analysis. As noted in the training protocol , we retained 97% of the signals that achieved an MSE below 0.4.

7. Results

A total of 5,528 significant SNPs with a threshold of $-\log_{10}(p) > 8.8$ were identified. After removing duplicates, **2,000 unique genes remained** for the STRING analysis.

(A) **(B)**

Figure 7: STRING network visualizations of enriched genes.

(A) Voltage-Gated Channel and Calcium-Activated Potassium Channel Activity - *Controlling the Heart's Electrical Signals*

This cluster focuses on genes that help control the movement of electrical signals in the heart, which are essential for keeping a regular heartbeat. It includes genes involved in:

- **Voltage-gated channels**: Ion channels that open or close in response to changes in electrical voltage.
- **Calcium-activated potassium channels**: These open when calcium levels inside the cell rise, and help reset the signal so it's ready for the next heartbeat.

Genes like **KCNQ5**, **KCND2**, and **CACNA1C** help manage these electrical signals. They are especially important for maintaining heart rhythm and making sure the heart beats properly.
Disruptions in the activity of genes within this cluster have been associated with irregular heartbeats (arrhythmias) and other electrical conduction issues in the heart.
In simple terms, this cluster controls the electrical system of the heart, helping it beat at the right pace and with the right rhythm.

(B) Myofibril Assembly - *Building the Heart's Muscle Fibers*

This cluster includes genes that take part in assembling the muscle fibers of the heart. These fibers, called myofibrils, are the tiny threads inside heart muscle cells that allow the heart to contract and pump blood.
Some of the key genes in this group (like **TTN**, **TNNC2**, **TMOD3**, and others) help organize and connect the small components inside muscle cells, making sure everything is built properly and functions smoothly.
Because these genes are tightly connected in the network, it suggests that they work together in a coordinated way to keep the heart muscle strong and functioning well.
If this process is disturbed, it can lead to heart muscle diseases or other problems with muscle function.
In simple terms, this cluster is responsible for building and maintaining the machinery that makes the heart muscle contract.

As part of the STRING analysis, we focused on three biological layers: Tissue Expression, Human Phenotype, and Molecular Function. From each table, we extracted all terms related to the heart and cardiovascular system. In addition, we selected the top non-cardiovascular terms based on their strength, those that stood out the most within their category.
We retained all entries with a **false discovery rate (FDR) below 0.01**, in order to keep only statistically robust and highly significant findings. This threshold was chosen to focus on the most reliable results.
Next, we retained all rows with a **strength value higher than 0.2**, since low-strength values suggest weak or less biologically meaningful enrichment, even if the FDR is low. This two-step filtering allowed us to focus on terms that are not only statistically significant but also biologically relevant.
Table 3 below, presents the final list of top terms that passed both criteria. Many of these terms are related to heart function, such as coronary atherosclerosis, arterial disorders, heart rate, and blood pressure.
Beyond the cardiovascular system, the table also highlights other important biological signals. For example, terms related to metabolism (like glucose metabolism and triacylglycerol measurement), immune function (anti-thyroglobulin antibody), and neurological processes (actin binding) appear among the strongest non-heart-related enrichments. These findings suggest that the latent ECG representation captures not only cardiac activity but also broader physiological pathways associated with metabolic, immune, and nervous system functions.

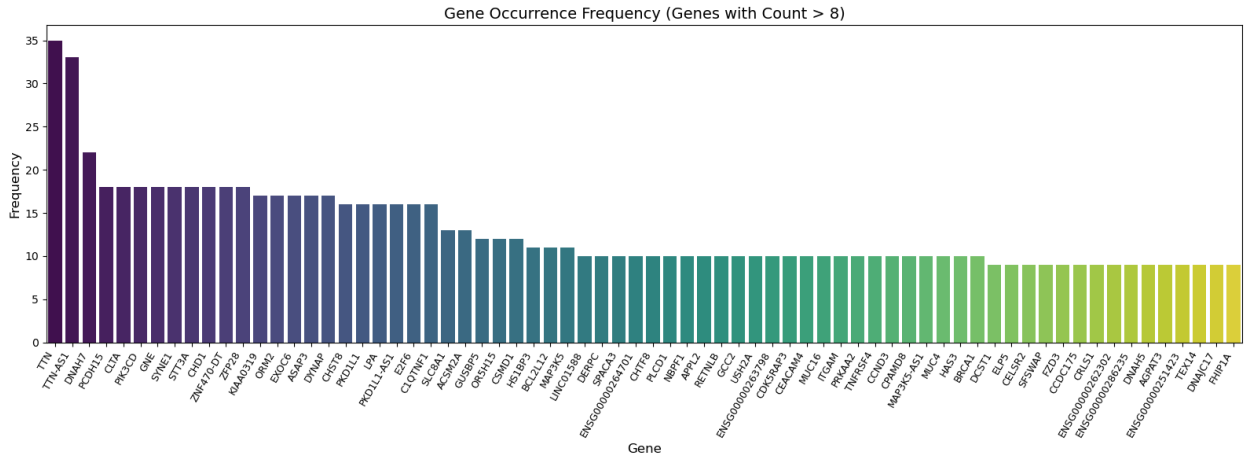
Term	Category	Strength	FDR
Coronary atherosclerosis measurement	**Cardiovascular**	**0.92**	**5.90e-03**
Arterial disorder	**Cardiovascular**	**0.87**	**2.90e-04**
Glucose metabolism measurement	Metabolic	0.77	4.40e-03
Triacylglycerol 52:2 measurement	Metabolic	0.76	7.20e-03
Susceptibility to cold sores measurement	Immune	0.76	6.00e-03
Anti-thyroglobulin antibody measurement	Immune	0.71	1.80e-03
Heart rate	**Cardiovascular**	**0.50**	**2.09e-10**
Phosphatidylcholine measurement	Metabolic	0.50	2.10e-04
Heart function measurement	**Cardiovascular**	**0.44**	**1.13e-28**
Cardiovascular disease	**Cardiovascular**	**0.42**	**2.55e-09**
Extracellular matrix structural constituent	**Cardiovascular**	**0.42**	**3.10e-03**
Heart disease	**Cardiovascular**	**0.41**	**2.20e-04**
Blood pressure	**Cardiovascular**	**0.37**	**2.99e-27**
Diastolic blood pressure	**Cardiovascular**	**0.36**	**3.57e-11**
Cardiovascular disease biomarker measurement	**Cardiovascular**	**0.33**	**1.31e-44**
Guanyl-nucleotide exchange factor activity	**Cardiovascular**	**0.33**	**5.20e-03**
Pulse pressure measurement	**Cardiovascular**	**0.30**	**2.75e-06**
Abnormal heart valve morphology	**Cardiovascular**	**0.28**	**3.40e-03**
Heart valve morphology	**Cardiovascular**	**0.28**	**3.40e-03**
Actin binding	Neurological	0.24	7.10e-03
Calcium ion binding	**Cardiovascular**	**0.23**	**5.60e-04**

Table 3: Top significant biological terms derived from the gene interaction network

The bar chart below shows the genes that appeared most often in our analysis (more than 8 times). Some genes clearly stand out, such as **TTN**, which appeared **35 times**. TTN is important for the **structure and flexibility of heart muscle**. Another frequent gene is **DNAH7** (**22 times**), which is involved in the movement of tiny cellular structures called cilia **and may influence how the heart develops and beats**.
Other genes with high counts are involved in different biological functions. For example:

- **PCDH15**, **SYNE1**, and **CHD1** help maintain **cell shape and communication**.
- **PIK3CD** is involved in **immune system signaling**.
- **GNE** is essential for **producing specific sugars used by cells**.

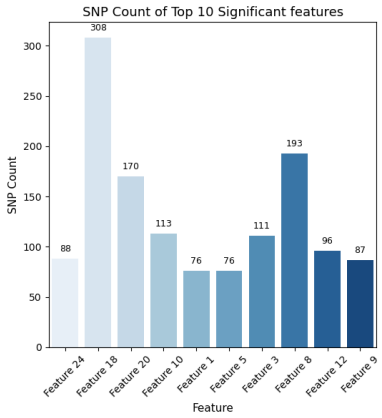
Some genes, like **LPA** (which affects **fat and cholesterol levels** and is linked to **heart disease risk**), and **SLC8A1** (which **controls calcium flow in heart cells**), directly support the **connection between genetics and heart function**.
Interestingly, many of the genes we found are **not limited to the heart**. Some are involved in **brain development, metabolism, or immune system activity** – such as **KIAA0319** (linked to **brain cell movement**), **MAP3K5**, and **ITGAM**.
This suggests that the **genetic patterns seen in ECG signals** may reflect not just **heart-specific processes**, but also **broader biological systems** that affect how the heart works.

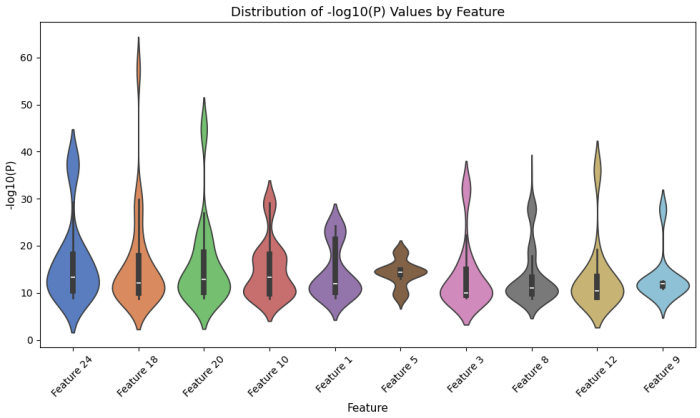


Distribution and Strength of SNP Associations by Latent ECG Features:

To identify the latent ECG features most influenced by genetic variation, we examined both the number and the distribution of genome-wide significant SNP associations using two complementary visualizations.
The bar plot on the left shows the number of genome-wide significant SNPs identified for each feature, representing the overall strength of associations with significant SNPs. Feature 18 and Feature 8 stand out with particularly high SNP counts (308 and 193 respectively), suggesting that these latent representations may capture physiologically relevant and heritable aspects of ECG morphology.
The violin plot on the right complements this by displaying the distribution of $-\log_{10}(p)$ values for the associated SNPs across the same features. While some features (e.g., Feature 24) show fewer associations but consistently high significance levels, others (e.g., Feature 20 and Feature 18) exhibit broader distributions, reflecting more variable association strengths.

Together, these plots reveal not only which features are genetically enriched, but also how concentrated or dispersed the association signals are. This helps prioritize features that may be driven by coherent biological mechanisms and guides downstream interpretation of the functional relevance of these ECG representations.

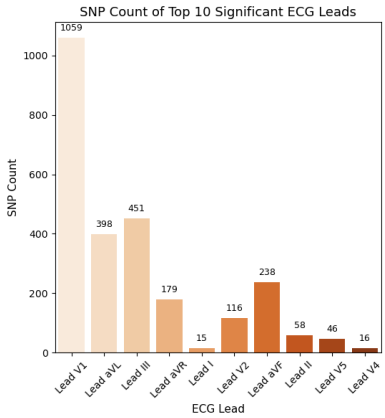


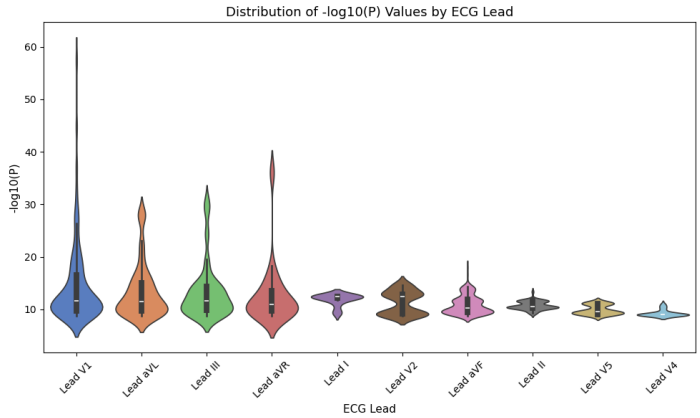


Distribution and Strength of SNP Associations by ECG Lead:

This figure presents two complementary views of genetic association patterns across standard ECG leads. Only the top 10 leads ranked by average $-\log_{10}(p)$ are shown, to highlight those with the strongest and most informative associations. The bar plot on the left shows, the total number of SNPs for a lead that reached genome-wide significance. The violin plot on the right illustrates the distribution of association strengths ($-\log_{10}(p)$) for those SNPs.

Lead V1 stands out in both number and strength of associations, with over 1,000 significant SNPs and a wide distribution of high $-\log_{10}(p)$ values. Leads aVL and III also show notable genetic signals, suggesting that they capture heritable aspects of cardiac physiology. In contrast, leads V4, V5, and II exhibit fewer and weaker associations, as seen by their lower bar heights and narrower violin shapes. This suggests that not all leads are equally influenced by genetic variation, and that septal and limb-oriented leads (e.g., V1, aVL) may be particularly informative in the context of genetic studies.





Gene–ECG Lead Association:

This is the central figure of our results - the heatmap that brings together the core of our analysis: the connection between genetic variation and ECG signals. It visualizes associations between genetic loci (X-axis) and ECG phenotypes, represented by different leads (Y-axis). Each cell reflects the number of SNPs significantly associated with a specific gene-phenotype pair.

The heatmap reveals notable variation in the number of genetic associations across different ECG leads. Lead V1 stands out as having the highest number of associated genes, suggesting a strong genetic influence on the electrical activity recorded from the right ventricle and the interventricular septum, which are the anatomical regions captured by this lead (placed on the 4th intercostal space at the right sternal border).

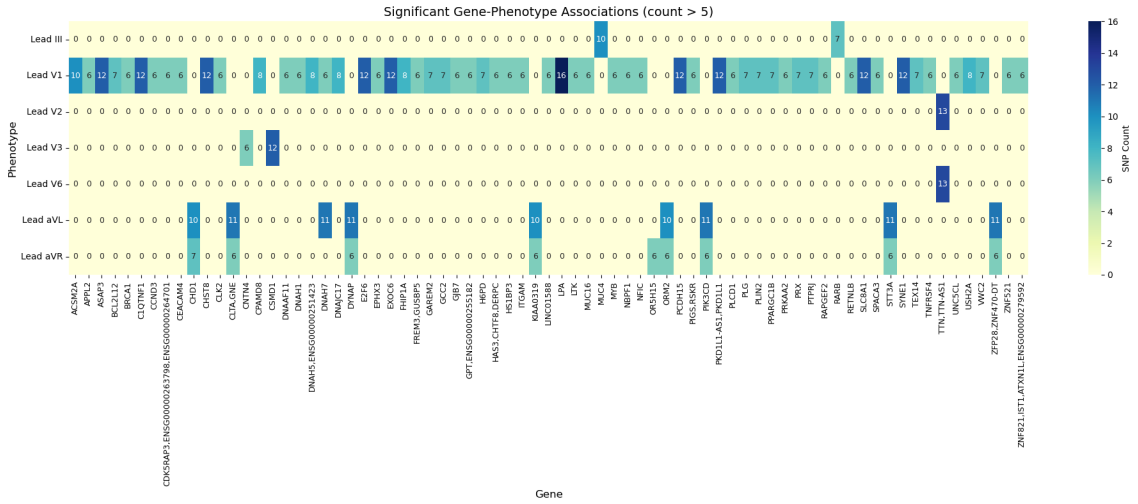
Among these genes, LPA shows the strongest connection to a Lead V1, with 16 significant genetic variations linked to it. The LPA gene is known for its role in fat and cholesterol processing, and is also associated with heart disease risk. Its strong link to the electrical signals in this part of the heart suggests that genes related to fat metabolism may also influence how the heart's electrical system works, possibly by affecting the heart's structure or energy use.

The anatomical explanation may not fully capture why Lead V1 dominates in genetic associations. In an ECG, Lead V1 is unique in its electrical waveform: it typically displays a small R wave followed by a deep S wave, making the QRS complex predominantly negative. This contrasts with leads placed more to the left of the chest (e.g., V5, V6), where the R wave is larger and the signal is mostly positive, reflecting the dominant electrical activity of the left ventricle. While one might expect the leads over the left ventricle to reveal stronger signals, it is possible that the distinct waveform of V1 enhances sensitivity to subtle genetic variations - perhaps because deviations in conduction timing or wave shape stand out more clearly in this lead's configuration. Thus, V1's dominance may result not only from its anatomical position, but also from its physiological "perspective" on cardiac electrical activity.

Following Lead V1, Lead aVL also shows many genetic associations. This lead records electrical activity from the upper left side of the heart, suggesting that genes may influence how electrical signals travel through that region. In addition, Lead aVR, which usually provides a kind of "mirror view" of the left side of the heart, also shows several associations. This is interesting because aVR is not often used in clinical interpretation, yet here it seems to reveal genetic influences that are often overlooked.

Interestingly, the heatmap includes several genes with well-established cardiac roles (e.g., TTN, SLC8A1, DNAH7), as well as genes with less obvious cardiac connections (e.g., MYB, ZNF521, APPL2), suggesting that ECG signals are shaped by a combination of direct cardiac genes and genes with broader systemic or regulatory functions.

These findings emphasize that genetic influences on ECG morphology are not evenly distributed across the heart, and that certain anatomical regions, especially those viewed by V1, aVL, and aVR, may be more genetically sensitive or reveal clearer signals of association.



The interpretation of the results was supported by artificial intelligence tool (ChatGPT), and validated against established biomedical resources such as GWAS Catalog, PubMed, WikiRefua, and Infomed to ensure reliability and biological relevance.

8. Conclusions

This study aimed to uncover **genetic factors that influence resting ECG morphology** by examining **genome-wide significant associations between SNPs and latent ECG features**. Our analysis identified **5,528 significant SNPs** ($-\log_{10}(p) > 8.8$), mapped to **2,000 unique genes** after filtering, revealing **multiple biologically meaningful patterns**.
Functional enrichment clustered the associated genes into two major biological domains:

1. **Cardiac electrical regulation**, including genes such as **KCNQ5** and **CACNA1C**, which are involved in **voltage-gated and calcium-activated potassium channels** essential for cardiac rhythm;

2. **Myofibril structure and contraction**, including **TTN** and **TMOD3**, which are responsible for building and maintaining heart muscle fibers.

Visualization of SNP distributions across latent ECG features highlighted **Features 18, 8, and 24** as **genetically enriched components**. While **Feature 18** had the largest number of significant SNPs, **Feature 24** showed more **consistently high association strengths**, suggesting different forms of **heritable contribution to ECG signal variability**.
Similarly, analysis across standard ECG leads revealed that **Leads V1 and aVL are especially informative genetically**. **Lead V1** exhibited **over 1,000 significant associations**, particularly with genes like **LPA**, **TTN**, and **DNAH7**, indicating a **strong connection between septal electrical activity and underlying genetic regulation**. These findings suggest that **not all ECG leads carry equal genetic relevance**, and that **septal and limb leads may serve as sensitive indicators of cardiac-genetic interplay**.
STRING-based functional annotation further revealed that while many enriched terms are **heart-specific** (e.g., **heart rate, atherosclerosis**), others point to **broader physiological processes**, such as **metabolism, immune response, and even neurological function**. This supports the idea that **ECG signals may integrate information from both cardiac-specific and systemic biological pathways**.
In conclusion, **resting ECG signals** – especially certain **latent components and leads** carry **strong and diverse genetic signatures**. These findings not only enhance our understanding of the **heritability of ECG traits** but also suggest **novel biological mechanisms linking cardiovascular function with systemic health**.

9. Future Directions and Challenges

While this study focused on resting ECG morphology, a key challenge for future research is to understand how genetic influences manifest under dynamic physiological states, such as during physical activity or stress. Investigating ECG signals during movement may reveal additional layers of genetic regulation that are not detectable at rest. Another major goal is to link specific genetic profiles to clinical phenotypes, including arrhythmias and cardiomyopathies, to enhance risk prediction and disease understanding. Finally, a promising direction is the development of models that connect gene expression patterns to physiological states—both resting and active—captured via ECG, paving the way for personalized cardiac monitoring based on genotype-driven signal interpretation.

10. Bibliography

[1] Ozal Yildirim, Ru San Tan, and U. Rajendra Acharya. An efficient com-pression of ecg signals using deep convolutional autoencoders. *Cognitive Systems Research*, 52:198–211, Dec 2018.

[2] Ling Zhou, Amanda K. Johnson, Cory M. Anderson, et al. Deep learn-ing–augmented ecg analysis for screening and genotype prediction of con-genital long qt syndrome. *JAMA Cardiology*, 9(1):65–74, 2024.

[3] Kimmo Porthan, Annukka Marjamaa, Matti Viitasalo, Heikki V¨a¨an¨anen, Antti Jula, Lauri Toivonen, Markku S. Nieminen, Christopher Newton-Cheh, Veikko Salomaa, Kimmo Kontula, and Lasse Oikarinen. Relationship of common candidate gene variants to electrocardiographic t-wave peak to t-wave end interval and t-wave morphology parameters. 7(7):898–903, Jul 2010. *Heart Rhythm*,

[4] Jessica van Setten, Niek Verweij, Hamdi Mbarek, Maartje N. Niemeijer, Stella Trompet, Dan E. Arking, Jennifer A. Brody, Ilaria Gandin, Niels Grarup, Leanne M. Hall, et al. Genome-wide association meta-analysis of 30,000 samples identifies seven novel loci for quantitative ecg traits. *European Journal of Human Genetics*, 27(6):952–962, Jun 2019.

[5] Lloyd T. Elliott, Kevin Sharp, Fidel Alfaro-Almagro, Sinan Shi, Karla L. Miller, Gwena¨elle Douaud, Jonathan Marchini, and Stephen M. Smith. Genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature*, 562(7726):210–216, Oct 2018.

[6] Ecg-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation*, 145(2):122–133, Jan 2022.

[7] Ewa Sieliwonczyk, Arunashis Sau, Konstantinos Patlatzoglou, Kathryn A. McGurk, Libor Pastika, Prisca K. Thami, Massimo Mangino, Sean L. Zheng, George Powell, Lara Curran, Rachel J. Buchan, Pantazis Theotokis, Nicholas S. Peters, Bart Loeys, Daniel B. Kramer, Jonathan W.Waks, Fu Siong Ng, and James S. Ware. Unsupervised feature extraction using deep learning empowers discovery of genetic determinants of the electrocardiogram. *medRxiv*, 2024.

[8] Taedong Yun, Justin Cosentino, Babak Behsaz, Zachary R. McCaw, Davin Hill, Robert Luben, Dongbing Lai, John Bates, Howard Yang, Tae- Hwi Schwantes-An, Yuchen Zhou, Anthony P. Khawaja, Andrew Carroll, Brian D. Hobbs, Michael H. Cho, Cory Y. McLean, and Farhad Hormoz- diari. Unsupervised representation learning improves genomic discovery and risk prediction for respiratory and circulatory functions and diseases. *medRxiv*, Preprint:2023.04.28.23289285, Aug 2023.

[9] Niek Verweij, Jan-Walter Benjamins, Michael P. Morley, Yordi J. van de Vegte, Alexander Teumer, Teresa Trenkwalder, Wibke Reinhard, Thomas P. Cappola, and Pim van der Harst. The genetic makeup of the electrocardiogram. *Cell Systems*, **11**(3):229–238.e5, Sep 2020.

[10] Dzemila Sero, Arslan Zaidi, Jiarui Li, Julie D. White, Tom´as B. Gonz´alez Zarzar, Mary L. Marazita, Seth M. Weinberg, Paul Suetens, Dirk Van-dermeulen, Jennifer K. Wagner, Mark D. Shriver, and Peter Claes. Facial recognition from dna using face-to-dna classifiers. *Nature Communications*, 10(1):2557, Jun 2019.

[11] William J. Young, Najim Lahrouchi, Aaron Isaacs, ThuyVy Duong, Luisa Foco, Farah Ahmed, Jennifer A. Brody, Reem Salman, Raymond Noordam, Jan-Walter Benjamins, et

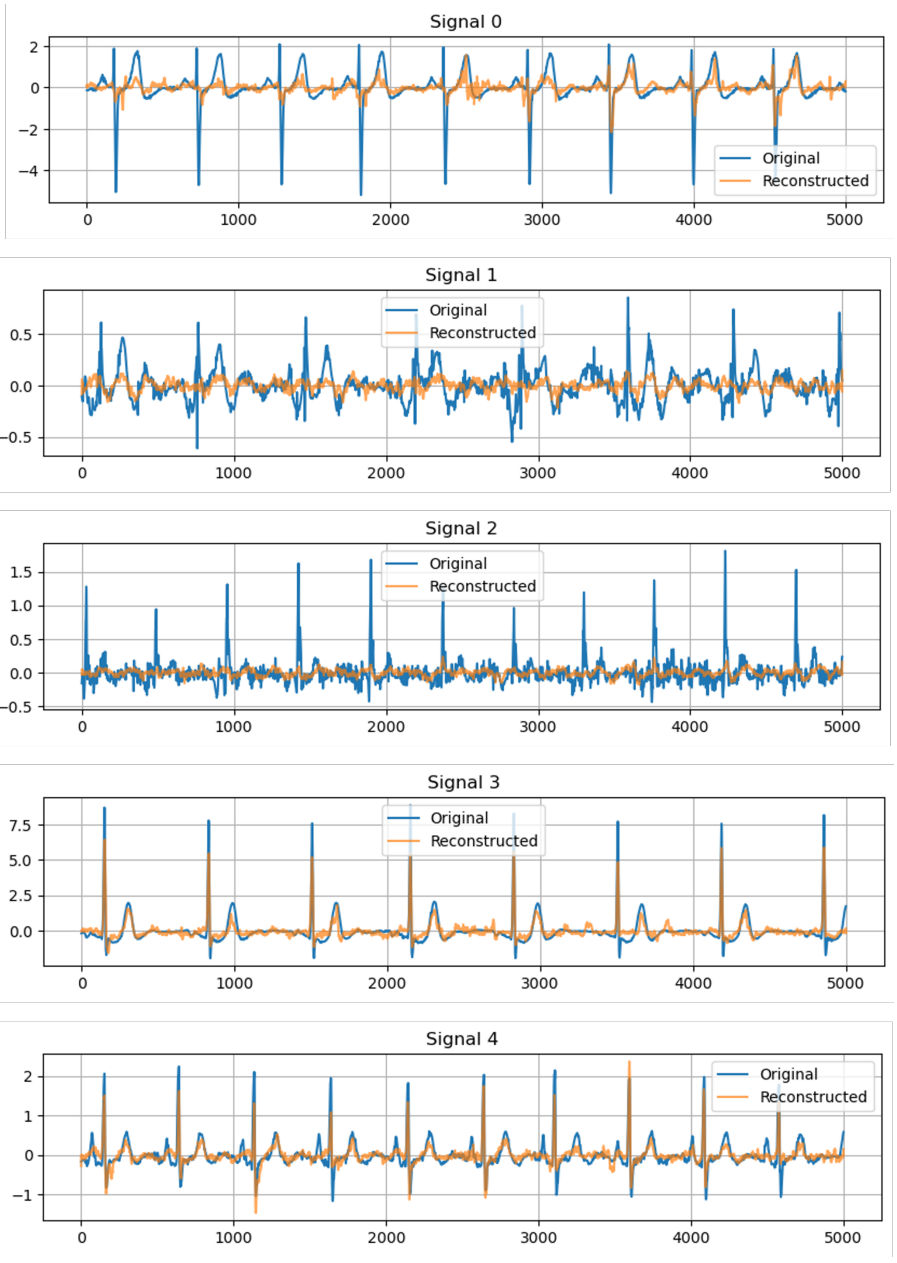
al. Genetic analyses of the electrocardiographic qt interval and its components identify additional loci and pathways. *Nature Communications*, 13(1):5144, Sep 2022.

[12] C. Matthew Sundling, Nagamani Sukumar, Hongmei Zhang, Curt M. Breneman, and Mark J. Embrechts. *Wavelets in Chemistry and Cheminformatics*. Rensselaer Polytechnic Institute, Troy, NY, USA.

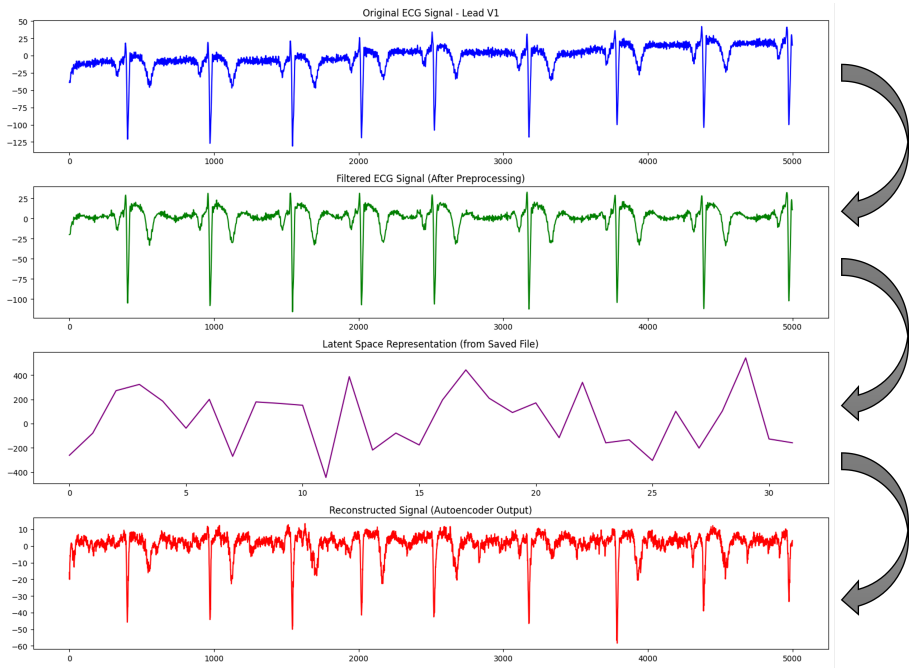
11. Appendices

GitHub Link: Click Here

Signal Reconstruction:



ECG Signal Process (from cleaning to reconstruction):



MSE Loss of All the ECG Signals:

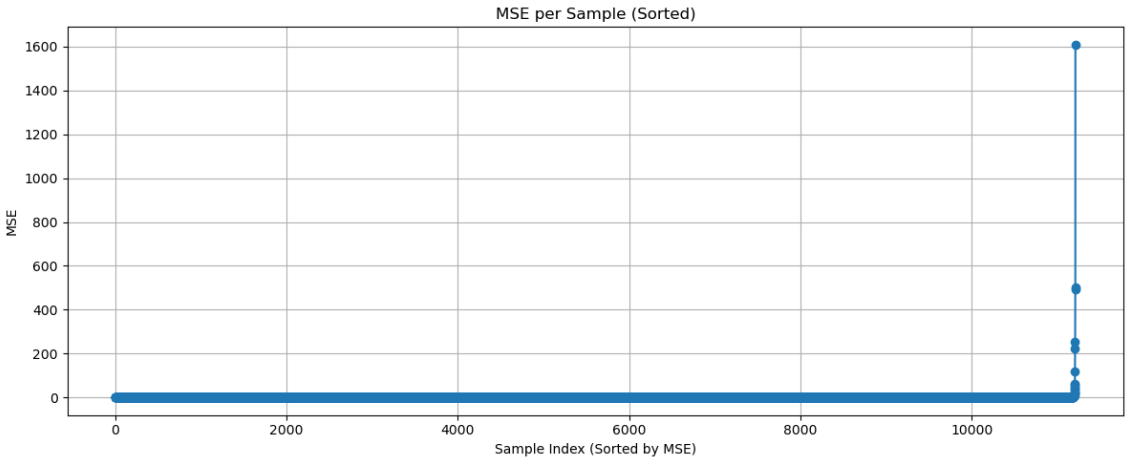
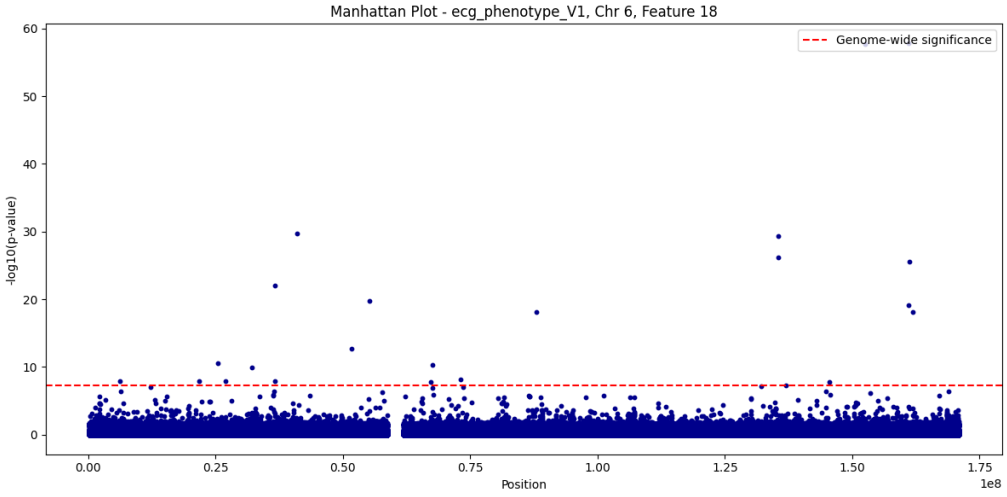


Table of ECG Signals with MSE Loss < 0.4:

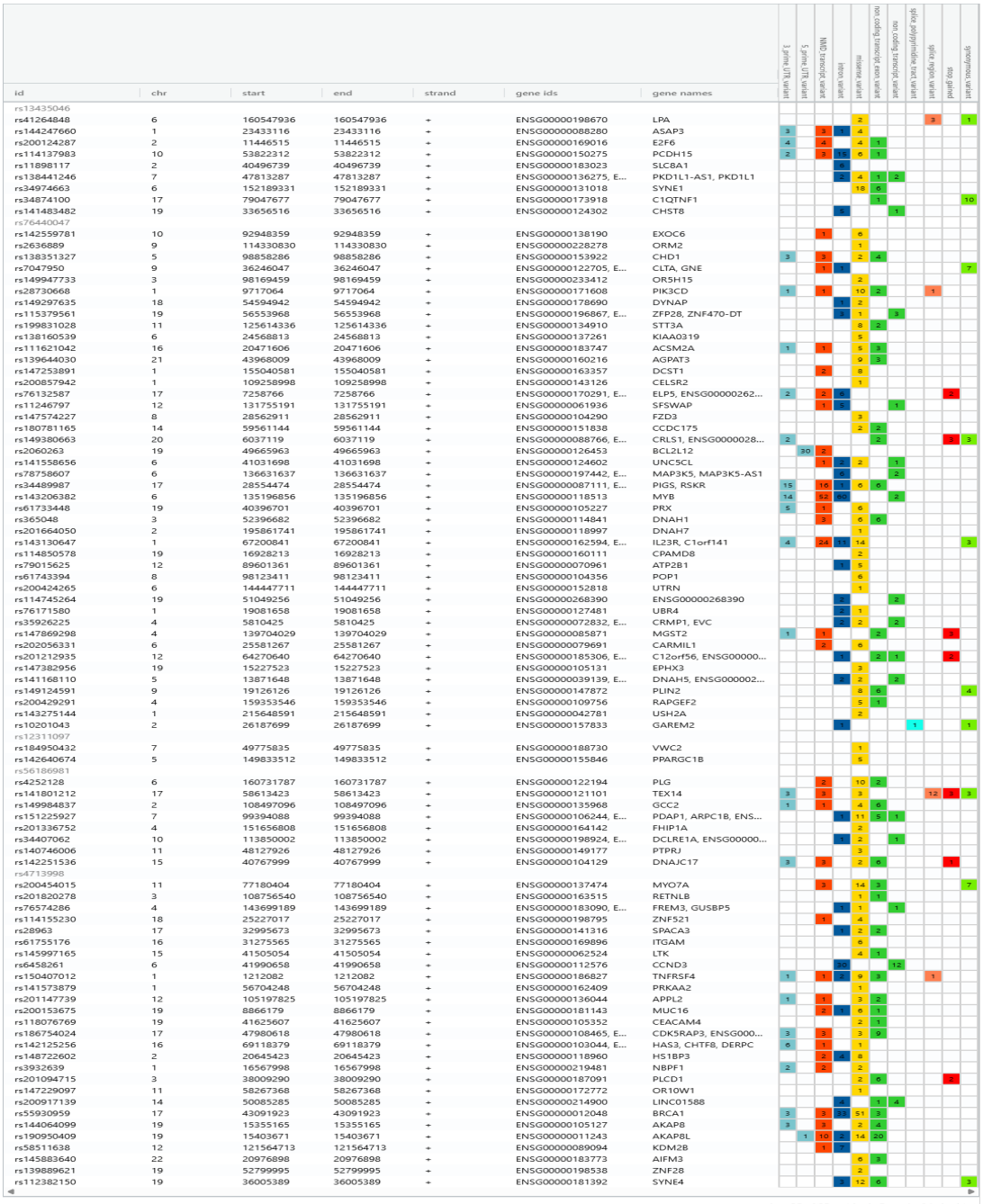
	mse	file_id	lead
0	0.001172	1031476	II
1	0.001309	1037716	II
2	0.001502	1075306	II
3	0.001535	1014066	II
4	0.001596	1092875	II
...	...	...	...
10778	0.397814	1002391	V6
10779	0.397880	1094162	V2
10780	0.398418	1005730	V6
10781	0.399015	1057915	V4
10782	0.399520	1019848	V2

10783 rows × 3 columns

Another Manhattan Plot Example Containing Significant SNPs:



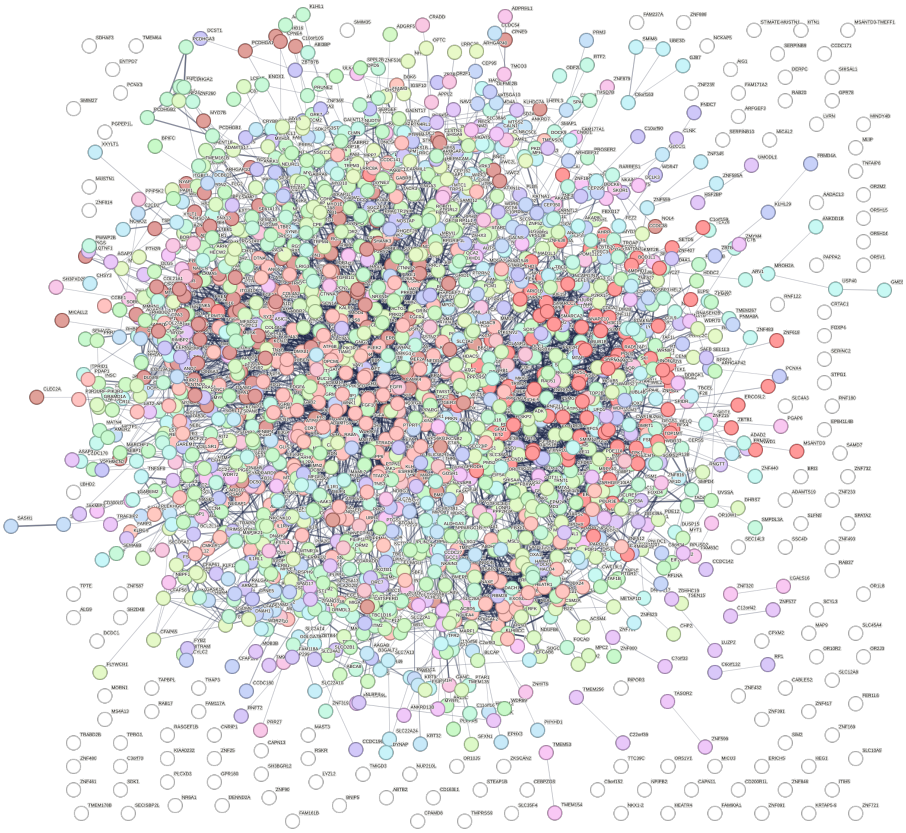
g:Profiler:



id	chr	start	end	strand	gene ids	gene names	3_prime_UTR_variant	5_prime_UTR_variant	NMD_transcript_variant	intron_variant	missense_variant	non_coding_transcript_exon_variant	non_coding_transcript_variant	splice_polypyrimidine_tract_variant	splice_region_variant	stop_gained	synonymous_variant
rs13435046																	
rs41264848	6	160547936	160547936	+	ENSG00000198670	LPA					2					3	1
rs144247660	1	23433116	23433116	+	ENSG00000088280	ASAP3	3		3	1	4						
rs200124287	2	11446515	11446515	+	ENSG00000169016	E2F6	4		4		4	1					
rs114137983	10	53822312	53822312	+	ENSG00000150275	PCDH15	2		3	15	6	1					
rs11898117	2	40496739	40496739	+	ENSG00000183023	SLC8A1				4							
rs138441246	7	47813287	47813287	+	ENSG00000136275, E...	PKD1L1-AS1, PKD1L1				2	4	1	2				
rs34974663	6	152189331	152189331	+	ENSG00000131018	SYNE1					18	6					
rs34874100	17	79047677	79047677	+	ENSG00000173918	C1QTNF1						1					10
rs141483482	19	33656516	33656516	+	ENSG00000124302	CHST8				2			1				
rs76440047																	
rs142559781	10	92948359	92948359	+	ENSG00000138190	EXOC6			1		6						
rs2636889	9	114330830	114330830	+	ENSG00000228278	ORM2					1						
rs138351327	5	98858286	98858286	+	ENSG00000153922	CHD1	3		3		2	4					
rs7047950	9	36246047	36246047	+	ENSG00000122705, E...	CLTA, GNE			1	1							7
rs149947733	3	98169459	98169459	+	ENSG00000233412	OR5H15					2						
rs28730668	1	9717064	9717064	+	ENSG00000171608	PIK3CD	1		1		10	2			1		
rs149297635	18	54594942	54594942	+	ENSG00000178690	DYNAP				1	2						
rs115379561	19	56553968	56553968	+	ENSG00000196867, E...	ZFP28, ZNF470-DT				2	1		3				
rs199831028	11	125614336	125614336	+	ENSG00000134910	STT3A					8	2					
rs13160539	6	24568813	24568813	+	ENSG00000137261	KIAA0319					5						
rs111621042	16	20471606	20471606	+	ENSG00000183747	ACSM2A	1		1		5	3					
rs139644030	21	43968009	43968009	+	ENSG00000160216	AGPAT3					9	3					
rs147253891	1	155040581	155040581	+	ENSG00000163357	DCST1			2		8						
rs200857942	1	109258998	109258998	+	ENSG00000143126	CELSR2					1						
rs76132587	17	7258766	7258766	+	ENSG00000170291, E...	ELP5, ENSG00000262...	2		2	6						2	
rs11246797	12	131755191	131755191	+	ENSG00000061936	SFSWAP			1	5			1				
rs147574227	8	28562911	28562911	+	ENSG00000104290	FZD3					3						
rs180781165	14	59561144	59561144	+	ENSG00000151838	CCDC175					2	2					
rs149380663	20	6037119	6037119	+	ENSG00000088766, E...	CRLS1, ENSG0000028...	2					2				2	3
rs2060263	19	49665963	49665963	+	ENSG00000126453	BCL2L12		30	2								
rs141558656	6	41031698	41031698	+	ENSG00000124602	UNC5CL			1	3	2		1				
rs78758607	6	136631637	136631637	+	ENSG00000197442, E...	MAP3K5, MAP3K5-AS1				5			2				
rs34489987	17	28554474	28554474	+	ENSG00000087111, E...	PIGS, RSKR	15		16	3	6	6					
rs143206382	6	135196856	135196856	+	ENSG00000118513	MYB	14		52	50			2				
rs61733448	19	40396701	40396701	+	ENSG00000105227	PRX	5		1		6						
rs365048	3	52396682	52396682	+	ENSG00000114841	DNAH1			3		6	6					
rs201664050	2	195861741	195861741	+	ENSG00000118997	DNAH7					1						
rs143130647	1	67200841	67200841	+	ENSG00000162594, E...	IL23R, C1orf141	4		24	13	14						3
rs114850578	19	16928213	16928213	+	ENSG00000160111	CPAMD8					2						
rs79015625	12	89601361	89601361	+	ENSG00000070961	ATP2B1				13	5						
rs61743394	8	98123411	98123411	+	ENSG00000104356	POP1					6						
rs200424265	6	144447711	144447711	+	ENSG00000152818	UTRN					1						
rs114745264	19	51049256	51049256	+	ENSG00000268390	ENSG00000268390				2			2				
rs76171580	1	19081658	19081658	+	ENSG00000127481	UBR4				3	1						
rs35926225	4	5810425	5810425	+	ENSG00000072832, E...	CRMP1, EVC				2	2		2				
rs147869298	4	139704029	139704029	+	ENSG00000085871	MGST2	1		1			2				3	
rs202056331	6	25581267	25581267	+	ENSG00000079691	CARMIL1			2		6						
rs201212935	12	64270640	64270640	+	ENSG00000185306, E...	C12orf56, ENSG00000...				11		2	1			2	
rs147382956	19	15227523	15227523	+	ENSG00000105131	EPHX3					3						
rs141168110	5	13871648	13871648	+	ENSG00000039139, E...	DNAH5, ENSG000002...				2	2		2				
rs149124591	9	19126126	19126126	+	ENSG00000147872	PLIN2					8	6					4
rs200429291	4	159353546	159353546	+	ENSG00000109756	RAPGEF2					5	1					
rs143275144	1	215648591	215648591	+	ENSG00000042781	USH2A					2						
rs10201043	2	26187699	26187699	+	ENSG00000157833	GAREM2				1				1			1
rs12311097																	
rs184950432	7	49775835	49775835	+	ENSG00000188730	VWC2					1						
rs142640674	5	149833512	149833512	+	ENSG00000155846	PPARGC1B					5						
rs56186981																	
rs4252128	6	160731787	160731787	+	ENSG00000122194	PLG			2		10	2					
rs141801212	17	58613423	58613423	+	ENSG00000121101	TEX14	3		3		3				12	3	3
rs149984837	2	108497096	108497096	+	ENSG00000135968	GCC2	1		1		4	6					
rs151225927	7	99394088	99394088	+	ENSG00000106244, E...	PDAP1, ARPC1B, ENS...				11	11	5	1				
rs201336752	4	151656808	151656808	+	ENSG00000164142	FHIP1A					2						
rs34407062	10	113850002	113850002	+	ENSG00000198924, E...	DCLRE1A, ENSG00000...				11	2		1				
rs140746006	11	48127926	48127926	+	ENSG00000149177	PTPRJ					3						
rs142251536	15	40767999	40767999	+	ENSG00000104129	DNAJC17	3		3		2	6				3	
rs4713998																	
rs200454015	11	77180404	77180404	+	ENSG00000137474	MYO7A			3		14	3					7
rs201820278	3	108756540	108756540	+	ENSG00000163515	RETNLB					1	1					
rs76574286	4	143699189	143699189	+	ENSG00000183090, E...	FREM3, GUSBP5				11	1		1				
rs114155230	18	25227017	25227017	+	ENSG00000198795	ZNF521			1		4						
rs28963	17	32995673	32995673	+	ENSG00000141316	SPACA3				11	2	2					
rs61755176	16	31275565	31275565	+	ENSG00000169896	ITGAM					6						
rs145997165	15	41505054	41505054	+	ENSG00000062524	LTK					4	1					
rs6458261	6	41990658	41990658	+	ENSG00000112576	CCND3				35			12				
rs150407012	1	1212082	1212082	+	ENSG00000186827	TNFRSF4	1		1	3	9	3			1		
rs141573879	1	56704248	56704248	+	ENSG00000162409	PRKAA2					1						
rs201147739	12	105197825	105197825	+	ENSG00000136044	APPL2	1		1		3	2					
rs200153675	19	8866179	8866179	+	ENSG00000181143	MUC16			2	1	6	1					
rs118076769	19	41625607	41625607	+	ENSG00000105352	CEACAM4					2	1					
rs186754024	17	47980618	47980618	+	ENSG00000108465, E...	CDK5RAP3, ENSG000...	3		3		3	9					
rs142125256	16	69118379	69118379	+	ENSG00000103044, E...	HAS3, CHTF8, DERPC	6		1		1						
rs148722602	2	20645423	20645423	+	ENSG00000118960	HS1BP3			2	11	8						
rs3932639	1	16567998	16567998	+	ENSG00000219481	NBPF1	2		2		2						
rs201094715	3	38009290	38009290	+	ENSG00000187091	PLCD1					2	6				2	
rs147229097	11	58267368	58267368	+	ENSG00000172772	OR10W1					1						
rs200917139	14	50085285	50085285	+	ENSG00000214900	LINC01588				4		1	4				
rs55930959	17	43091923	43091923	+	ENSG00000012048	BRCA1	3		3	33	51	3					
rs144064099	19	15355165	15355165	+	ENSG00000105127	AKAP8	3		3		2	4					
rs190950409	19	15403671	15403671	+	ENSG00000011243	AKAP8L		1	10	2	14	20					
rs58511638	12	121564713	121564713	+	ENSG00000089094	KDM2B			1	7							
rs145883640	22	20976898	20976898	+	ENSG00000183773	AIFM3					6	3					
rs139889621	19	52799995	52799995	+	ENSG00000198538	ZNF28					2						
rs112382150	19	36005389	36005389	+	ENSG00000181392	SYNE4				3	12	6					3

1 to 95 of 95 | Page 1 of 1

STRING – Full Protein-Protein Interaction Network:



42 Clusters Out of 338:

color	cluster Id	gene count	description
	Cluster 1	94	DNA repair
	Cluster 2	86	+ Transmembrane receptor protein tyrosine kinase signaling pathway
	Cluster 3	49	+ ECM-receptor interaction
	Cluster 4	39	+ G alpha (q) signalling events
	Cluster 5	36	+ Protein-protein interactions at synapses
	Cluster 6	31	Preribosome, and Ribosome biogenesis
	Cluster 7	18	+ Sensory perception of sound
	Cluster 8	17	+ Myofibril assembly
	Cluster 9	17	Voltage-gated channel, and Calcium-activated potassium channel activity
	Cluster 10	16	Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases
	Cluster 11	15	GAIN domain superfamily, and YD repeat
	Cluster 12	14	Regulation of G protein-coupled receptor signaling pathway
	Cluster 13	14	+ Cilium Assembly
	Cluster 14	13	RAC3 GTPase cycle
	Cluster 15	13	U2-type spliceosomal complex
	Cluster 16	13	+ Cilium movement
	Cluster 17	13	+ tRNA aminoacylation for protein translation
	Cluster 18	13	+ RHO GTPases Activate WASPs and WAVEs
	Cluster 19	12	+ Butanoate metabolism
	Cluster 20	11	Catenin complex
	Cluster 21	11	+ mRNA surveillance pathway
	Cluster 22	11	Purine metabolism
	Cluster 23	10	Interaction between L1 and Ankyrins
	Cluster 24	10	+ Sphingolipid de novo biosynthesis
	Cluster 25	10	Keratinization
	Cluster 26	10	Complex I biogenesis
	Cluster 27	9	Mixed, incl. MAP kinase activity, and Protein tyrosine/threonine phosphatase activity
	Cluster 28	9	COMM domain, and Protein neddylation
	Cluster 29	9	+ Kinesin complex
	Cluster 30	9	+ Negative regulation of viral genome replication
	Cluster 31	9	DAB1, MAGI1, MAGI2, RAPGEF2, RAPGEF6, SAMD12, STARD7, SYNPO, TRIM3
	Cluster 32	9	+ Negative regulation of muscle cell differentiation
	Cluster 33	9	Mixed, incl. Inter-alpha-trypsin inhibitor heavy chain C-terminus, and Alpha-1-acid gl...
	Cluster 34	9	+ mRNA surveillance pathway
	Cluster 35	9	O-linked glycosylation of mucins
	Cluster 36	8	+ Tandem pore domain potassium channels
	Cluster 37	8	Mixed, incl. ATAXIN1-like, and Ubiquilin
	Cluster 38	8	Signaling by ALK fusions and activated point mutants
	Cluster 39	8	Vision
	Cluster 40	8	+ Arachidonic acid metabolism
	Cluster 41	8	Xenobiotic transport
	Cluster 42	8	Regulation of commissural axon pathfinding by SLIT and ROBO