Implications of the Associations Between Structural Variants and Single Nucleotide Polymorphisms for Coronary Artery Disease Risk

Boris Pavić

Supervisors: Marcel Reinders, Niccolò Tesi

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Boris Pavić
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Niccolò Tesi, Andy Zaidman

Abstract

Coronary artery disease (CAD) is a condition characterized by the narrowing or blockage of the arteries that supply blood to the heart. It is a major global health burden and is known to be correlated with genetics, but the details of the genetic contribution remain unclear. In this study, we explored the associations between structural variants (SVs) and single nucleotide polymorphisms (SNPs) identified through previous CAD genome-wide association studies (GWAS). We identified the most relevant associations by ranking them using a proposed composite score, combining effect sizes and p-values. Filtering was applied to keep the most significant associations within a 500kb genomic window, resulting in 968 SNP-SV associations. Cross-referencing with tissue-specific eQTL data from the GTEx Portal indicated 321 SNP-SV associations that impact gene expression in coronary artery tissue. GSEA identified associated pathways involving dopachrome isomerase, phenylpyruvate tautomerase, and catalytic activity. The results suggest that the SVs make an important contribution to the regulation of CAD-related gene expression and the overall risk of CAD.

1 Introduction

Despite significant advancements in modern-day medicine, cardiovascular diseases continue to be a major cause of death throughout the world. The most prevalent of these disorders is coronary artery disease (CAD); a condition characterized by the narrowing or blockage of the arteries that supply blood to the heart, known as coronary arteries [1]. Unlike many other diseases, symptoms of this disease occur at a stage at which it is already far advanced, making it particularly dangerous [2]. Being aware of the personal risk of developing CAD is thus crucial for maintaining one's well-being, as it motivates regular health checkups through which the disease progress, if any, can be tracked and acted upon before it's too late. While many of the known risk factors can be controlled, such as diet choices and physical activity, it is known that genetics also have a significant contribution to CAD risk [3]. Still, the details of this contribution remain unclear.

The carrier of genetic information of an organism is deoxyribonucleic acid, the molecule better known as DNA. It comes in the form of a double helix, being a sequence of pairs of nucleotides, the so-called building blocks of DNA. We distinguish four types of nucleotides in DNA, based on their nucleobases: adenine (A), thymine (T), cytosine (C), and guanine (G). The pairs forming the sequence are always either adenine-thymine or cytosine-guanine. Segments of this sequence form genes that contain the information needed to produce proteins, which influence biological processes and affect human traits. The exact structures of these genes aren't universal across our species but instead vary in the form of alleles. A common example used to illustrate this is eye color, as the gene that determines it has different alleles (forms) for brown and blue eyes. Such alleles are often determined by genetic variations - changes in the nucleotide sequence of DNA. These variations have downstream effects on gene expression and protein function.

Studying these genetic variations is crucial to our understanding of the development of various traits and diseases. Identifying these variations enables pinpointing the parts of the genome that show an association with the studied trait and hypothesizing their upstream effect. A prominent approach to studying the genetic associations in this manner is through genome-wide association studies (GWAS). GWAS identify differences in allele frequency of genetic variants between individuals with a particular trait of interest and a control group, revealing associations of these differences and the trait [4]. The most common and prevalent genetic variants studied in GWAS are single-nucleotide polymorphisms (SNPs). As their name suggests, SNPs indicate a change in a single nucleotide at a given position in the DNA (for example, A to G). These variants have already been extensively studied through many GWAS due to their simplicity and abundance [4]. As a result, we now have a greater understanding of the genetic underpinnings of many human traits, and many associations have already been identified. However, understanding the downstream consequences of the identified SNPs is often complicated.

To better understand the functional consequences of SNPs, quantitative trait loci (QTL) analyses are often used to link SNPs to quantitative traits. In this way, they identify the correlation of the SNPs with certain cellular functions for a particular tissue or cell type. This makes them a beneficial extension to the results obtained from GWAS, uncovering the low-level biological effects of the discovered trait-associated regions [5]. While this doesn't give the full picture of the trait's development, understanding the basic changes serves as an essential starting point in this process. The types of QTLs are categorized by the biological effect they embody. For example, the effects on gene expression are captured by eQTLs.

A more complex variations found in the genome are structural variants (SVs). They represent a genomic alteration involving a DNA segment larger than SNPs and are more precisely classified into types based on the nature of the structural change at play [6]. One such type of structural variants are tandem repeats, where a particular string of two or more nucleotides (also referred to as the repeat unit) is repeated adjacently a given number of times. One can imagine how this clear difference in structure complexity makes studying SVs considerably more challenging than SNPs. And to no surprise, the sequencing methods used for the genotyping of individuals confirm this suspicion. Whole-genome sequencing (WGS) is considered the most powerful sequencing method, being able to determine nearly the entire DNA sequence of a full genome, but despite its high potential it still fails to accurately capture some SVs [1]. Studying the associations between SVs and diseases through GWAS continues to be difficult as a result.

Interestingly, it has been hypothesized that the combined effect of SVs on trait expression is likely to be stronger than that of SNPs. While the exact effects are not well understood, they are known to have regulatory effects on gene expression in particular, and despite being less frequent than SNPs, their larger size results in larger nucleotide sequence differences and is likely to be the causal factor for this phenomenon [7]. This significant regulatory function has culminated in a growing interest in SVs, but progress is held back by technological limitations, thus requiring a change in perspective. This gave rise to the idea of studying the associations between SVs and traits indirectly. The newly established association dataset between SNPs and SVs in the human genome constitutes the basis for this idea, as it opens doors to investigating SVs through well-established SNPs by proxy. Given that many SNPs associated with CAD have already been identified by previous GWAS, such as the study by Van der Harst et al. in 2018 [8], it is possible to find such SVs that show significant association with these SNPs and thus likely contribute to CAD in their own right.

In this report, we aim to investigate what SVs are most significantly associated with CAD-SNPs and whether the discovered associations help explain the biological underpinnings of CAD. Additionally, we will inspect the relative eQTLs, the affected genes, and the biological processes they are associated with.

We started by establishing a dataset containing all SVs showing significant association with CAD-SNPs by combining the results from GWAS [8] with the SNP-SV association dataset. The obtained associations were then ranked based on a derived composite score, which combined the significance and effect sizes of SNP-SV and CAD-SNP associations, prioritizing the most significant ones. We then filtered the dataset by removing cluttered associations based on genomic position to focus on only the most significant ones per region. The filtered set of associations was then cross-matched against tissue-specific eQTL data from the GTEx Portal [9] to obtain a direct biological interpretation for gene expression of the associations deemed most significant. Finally, gene set enrichment analysis [10] was performed on the set of affected genes using g:Profiler [11] to determine the affected pathways. An overview of these steps is outlined in Fig 1.

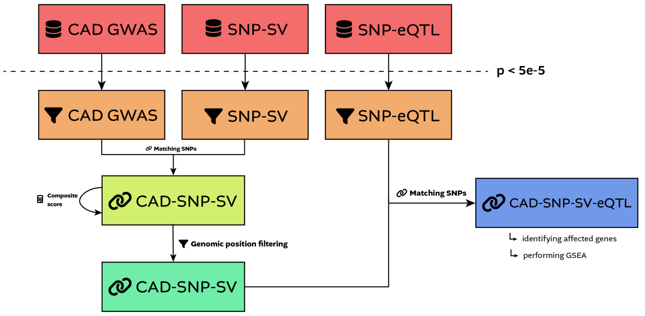


Figure 1: Overview of the methodology procedure. The datasets were first filtered by significance, keeping those with $p < 5e-5$. The filtered CAD GWAS and SNP-SV datasets were then combined based on matching SNPs. The associations were ranked by the computed composite score, and genomic position filtering was applied. The final set of associations was obtained by matching the SNPs with the filtered SNP-eQTL dataset.

2 Materials and Methods

2.1 Data sources

CAD GWAS

The CAD GWAS data originates from a study by Van der Harst et al. in 2018 [8]. The study included 34,541 CAD cases and 261,984 controls from the UK Biobank. The data includes information about the SNP position in the genome, the affected alleles, a beta value indicating the effect size of the SNP for CAD, as well as the p-value indicating the significance of the association. The data was downloaded from the European Bioinformatics Institute (EBI) FTP server[1].

SNP-SV associations

The SNP-SV association data comes from our own dataset. It was generated from a study that included 93 patients diagnosed with Alzheimer's disease from the Amsterdam University Medical Center [12] and 121 cognitively healthy centenarians from the 100-plus Study cohort [13]. The associations were derived through QTL analysis with PLINK [14], using linear regression models to link the SNPs to SV size (which is the quantitative trait in this scenario). In this step, they considered all SNPs within a 500kb up/downstream window from the SV location. Furthermore, the structural variants were classified as either tandem repeats (TRs), which were described previously, or transposable elements (TEs). TEs are segments of ancient bacterial DNA integrated into our genome, characterized by their dynamic nature of moving around the genome. This leads to them either being present or absent at a given position, which is why we characterize them as deletions and insertions (INDELs) respectively. This dataset contains similar information, such as SNP position, affected alleles, and the effect size of the SNP in regards to the SV size along with a corresponding p-value, the structural variant position, and type.

SNP-eQTL data

Finally, the eQTL data was obtained from the Adult GTEx open-access datasets at GTEx Portal [15], particularly the GTEx Analysis V8 cis-eQTLs mapped in European-American subjects for the coronary artery tissue. Notably, while other heart-related tissues could also have been considered, we decided to focus only on the coronary artery tissue, being the most relevant choice for CAD. Other than the already described information relating to the variant being matched, the dataset also includes the ID of the gene whose expression is affected along with the distance of the variant to the transcription start site (TSS) of the gene.

2.2 Initial Filtering

Each of the datasets had to undergo some initial filtering to reduce the information load and remove unnecessary pieces of data. In the case of the GWAS and SNP-SV data, this included removing all associations having $p > 5 \times 10^{-5}$. This is the significance threshold used by the GWAS catalog [16] as the inclusion criteria for new associations and was therefore considered a suitable cut-off point for this project.

[1] https://ftp.ebi.ac.uk/pub/databases/gwas/summary$_statistics/GCST005001-GCST006000/GCST005194/harmonised/$

The eQTL dataset for the coronary artery tissue contained data for genetic variants in general, but most of them were SNPs. For the sake of simplicity and ease of linking with the other datasets, it was filtered to only include SNP-eQTL associations while the other variant associations were removed. It was then filtered for significance, removing the associations having $p > 5 \times 10^{-5}$.

2.3 Combining the GWAS and SNP-SV data

The main idea behind this project lies in combining the CAD SNP and SNP-SV data in order to study the role of SVs in these associations. This was done in two steps. Firstly, the two datasets were joined on matching SNPs, such that the chromosome and position of the SNP is the same for both datasets. The second step required aligning the alleles of the two datasets to ensure data consistency. Thus, we first checked that the effect and alternative alleles of the two studies match to ensure that the same SNP is being considered. For example, a SNP at a given position should have A and G as either the effect or alternative allele in both of the datasets for it to be considered the same SNP. Then, the actual alignment is done. We check if the effect alleles of both studies are the same allele, and if not we flip the beta (effect size) of the SNP-SV association by multiplying it by -1. This essentially inverts the interpretation of the effect size by changing its direction, which is the same as considering the alternative allele as the effect allele. Once the alleles are aligned, we check that the beta from the GWAS dataset is positive, indicating that the given SNP increases the risk of CAD, and if not we flip both the GWAS and the SNP-SV beta value. This follows the same principle as the earlier inversion and is possible as the betas are already aligned prior to the calculation.

Before continuing with the next steps, we standardized the respective beta values with respect to the means and standard deviations of the values from the original datasets. We also applied correction for multiple testing to the individual p-values using the Bonferroni correction method [17], in an attempt to reduce their impact on the composite score, which is discussed next. This was done by multiplying the p-values with the number of tests, that being the number of associations obtained from combining the two datasets (38,665).

Composite score

Considering the large number of resulting associations, it was necessary to create a way to rank them to establish the most relevant ones. This was done in the form of a composite score, with the idea of accounting for the effect sizes and the significance values from both of the original datasets:

$$\text{Composite Score} = |\beta_{\text{std, CAD}}| + |\beta_{\text{std, SNP\_SV}}| + \log\left(\frac{1}{p_{\text{combined}}}\right)$$

, where $\beta_{\text{std, CAD}}$ and $\beta_{\text{std, SNP\_SV}}$ are the standardized betas for the CAD and SNP-SV datasets, respectively, and p_{combined} is the combined p-value. The combined p-value was calculated using the Fisher's method [18]:

$$\chi^2 = -2\left(\ln(p_{\text{CAD}}) + \ln(p_{\text{SNP\_SV}})\right)$$

, where χ^2 follows a chi-squared distribution with four degrees of freedom. The combined p-value is then obtained from the cumulative distribution function of the established chi-squared distribution.

Defining this composite score required several assumptions and considerations. The first of these assumptions is crucial for the use of Fisher's method, that being that the p-values of the two studies are independent. While it is impossible to prove the complete independence of these two values, considering that the datasets focus on different biological mechanisms and are obtained from separate, unrelated studies, it was deemed reasonable to consider them independent for the purposes of calculating the composite score. Another assumption is that standardizing the respective betas makes them comparable, despite coming from different studies. In other words, the two datasets are presumed to follow a similar beta distribution, and standardizing these values with their respective means and standard deviations results in a comparable construct. A similar assumption is made with regard to the independence of the two beta values, allowing for a direct summation of the two. It was also assumed that the direction of the combined effect is irrelevant, so the sum of absolute values can be used in the calculation. We essentially prioritize the magnitude of the effect size, rather than its direction, and since we previously ensured $\beta_{\text{std, CAD}}$ was positive, the only direction-changing factor is $\beta_{\text{std, SNP\_SV}}$, which carries less importance in the context of studying CAD. Finally, taking the logarithm of the combined p-value means highly significant associations end up with a high-magnitude term for their composite score. Because this term is negative by definition, it needs to be multiplied by -1 to lead to an increase in the overall score. When visualizing the composite score calculation, a sum of positive terms was deemed more intuitive. For this reason, the negative logarithm term was transformed to use the inverse of the p-value instead, leaving only positive terms in the equation.

It is important to note that the composite score serves as merely a ranking measurement, and despite these assumptions taking away from its mathematical correctness, it was still considered sufficient for the purposes of this research.

2.4 Genomic position filtering

With the composite score computed, the associations were ranked in descending order to identify those most relevant. However, the established list of combined associations in its raw form was too large to draw meaningful conclusions. Instead, another filtering step was taken to reduce the number of associations, while also isolating the most promising candidates. The algorithm starts by taking the first CAD-SNP-SV association from the list and calculating a genomic window 250kb up/downstream from the SNP position. It then iterates through all other associations looking for those within the defined window, with a matching structural variant to the selected association, and removes them. The process is repeated until all of the associations are considered, resulting in a list of the most promising genomically isolated SNPs and associated SVs. This filtering method is supported by the high correlation exhibited by SNPs close to one another. The correlated SNPs associate with the same SV, leading to redundant signals, so singling out those that are most relevant is more insightful. The pseudo-code depicting this algorithm in detail can be found in Appendix A.

It is important to note that due to the complex nature of SVs and the challenges of representing them, there are instances of the same SV being annotated as multiple different ones in the SNP-SV dataset. For example, a TE containing a segment that is also a TR will be labeled as two different SVs. To keep the matching and filtering simple, we considered such instances as separate SVs.

2.5 Cross-Matching with eQTL Data

The CAD-SNP-SV associations from the resulting dataset were cross-matched against the SNP-eQTL data to identify whether they affect gene expression in a meaningful way. This is done in a similar fashion to how the GWAS and SNP-SV data were combined, where we match the genomic position of the SNP between the two datasets and then align the alleles, flipping the eQTL beta if necessary. The resulting list of associations contains all of the previously described data along with eQTL-specific information, such as the gene affected and the effect size of the expression alteration.

2.6 Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) helps interpret the gene expression data by taking a set of genes, and checking whether the members of that set tend to be towards the top or bottom of the predetermined gene expression ranking list. In the case of this happening, the gene set is considered to be correlated to the trait being studied [10]. In other words, the genes from the set being mostly on the top (more active) or bottom (less active) indicates a change in their behavior for the studied trait.

The eQTL data contained the IDs of the genes being affected, but not the actual gene names. To circumvent this, the eQTL calculator provided by GTEx Portal [15] was used, as it takes a list of SNP IDs, gene IDs, and tissue names and returns eQTL data that also has the gene name. With that complete, a list of unique genes was identified, whose expression is presumably altered by the accompanying SNPs and SVs. GSEA was then performed on this set of genes using g:Profiler [11]. All of the provided background sets were used: Gene Ontology, biological pathways (KEGG, Reactome, WikiPathways), TRANSFAC, miRTarBase, HPA, CORUM, and Human Phenotype Ontology. The analysis included all gene sets from these background sets.

3 Results

3.1 Significant Associations Identified

The initial filtering of the three datasets resulted in 7,107,863 SNP-SV associations, 25,246 CAD-SNP associations, and 546,634 SNP-eQTL associations. This greatly reduced the load of information, allowing for significantly faster executions of the ensuing operations on the data.

After matching the SNPs between the CAD-SNP and SNP-SV datasets and aligning their alleles, 38,665 aligned CAD-SNP-SV associations were obtained. This was followed by ranking these associations with respect to the newly computed composite score. After filtering these associations based on genomic positions, 968 independent associations were left. This included 451 TRs and 517 TEs. An analysis of the associations revealed an equal frequency of associations with positive and negative effect sizes, both involving 484 associations. The means of the two sets were calculated, being $\mu^{+} = 367.206$ for the positive and $\mu^{-} = -257.113$ for the negative effect size set. This effect size indicates whether a SNP associates with longer (positive beta) or shorter (negative beta) SVs.

In order to decide on a fixed number of most significant associations to report, the distribution of the composite scores was plotted, as can be seen in Fig 2.

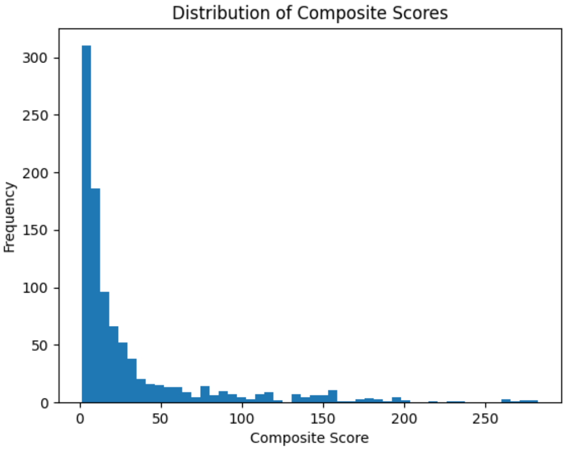


Figure 2: Distribution of Composite Scores. The plot shows the frequency of different composite scores among the 968 CAD-SNP-SV associations.

The plot indicates that the vast majority of the associations have a composite score lower than 75. This is followed by a low frequency of scores in the 75-175 range, which decreased even further after that point. For the purposes of this report, 175 was chosen as the cutoff threshold for the composite score, leaving a concise list of 27 associations shown in Table 1. The full list of associations is available in Appendix B.

The data indicates that the majority of the most highly correlated structural variants were transposable elements (16/27), whereas tandem repeats were less prevalent (11/27). These frequencies are similar to those observed in the full dataset. The highest significance for CAD is shown by SNPs rs10811647 and rs10811650, associated with SVs chr9:22057960-22058115_JOIN (TR) and chr9_22057962_D_2B1M8 (TE). The other associations show substantially lower CAD significance in comparison. A trend can be seen with the remaining associations having an extremely low SNP-SV p-value, which is the main contributor to their high composite scores.

12 associations have a positive and 15 associations have a negative SNP-SV effect size, roughly emulating the full dataset frequencies. Three of the 12 positive associations have effect size magnitudes larger than μ^{+}. Seven of the 15 negative associations have an effect size with a magnitude

larger than that of μ^-.

A notable inclusion is the column for the number of removed associations (NRA), referring to the number of associations that were removed during the genomic position filtering step for each association. Summing these values per SV type results in a combined NRA of 23,316 for TEs and 14,381 for TRs. A similar distribution is reflected in Table 1, with a summed NRA of 2,694 for TEs and 961 for TRs.

3.2 Linking the Associations with eQTLs

Cross-matching the established 968 CAD-SNP-SV associations with the SNP-eQTL dataset resulted in 321 associations. To get a better grasp of the relevance of these results, three volcano plots were made to show the relationship between the standardized effect sizes and p-values with respect to each of the original datasets. These plots are shown in Fig 4. Rather than highlighting the associations based on the direction of effect size, we opted for highlighting 10 SNPs showing the highest significance in their association with CAD. Unlike the associations from Table 1, these 10 SNPs were taken from this new result set of associations that includes the eQTLs.

The highlighted points in Fig 4a are also the most significant ones, as that was the criteria for their selection in the first place. All of the points have a positive effect size for CAD, as was ensured during the aligning process, and are well above the 5e-05 p-value threshold.

Fig 4b shows the same 10 SNPs but in relation to SVs. These associations are less significant than some of the other, non-highlighted points in the plot. Despite this, they still show high significance and are all above the p-value threshold. Eight of the associations show a positive, and two show a negative effect size.

Finally, Fig 4c shows the 10 SNPs in relation to eQTLs. The associations yet again show high significance and surpass the p-value threshold. Four of the associations have a positive and six have a negative effect size.

When comparing the plots, some SNPs stand out as being highly significant across all three datasets. For example, SNP rs6728861 (orange) significantly increases CAD risk ($p =$ 1.291e-32, Fig 4a). This SNP is significantly associated with a TE (chr2_203034349_D_19E0M1, $p =$ 1.07239e-75, Fig 4b). The SNP is also an eQTL for the CEP63 gene in coronary artery disease ($p =$ 3.21363e-14, Fig 4c) and is associated with increased expression of this gene. Considering the effect sizes, this indicates that a longer TE might increase the expression of CEP63.

Similarly, SNP rs8046696 (cyan) also significantly increases the risk of CAD ($p =$ 1.905e-16, Fig 4a). It shows a significant association with a TE (chr16_75395934_D_B8FMF, $p =$ 2.55858e-131, Fig 4b). This SNP is also an eQTL for the RBM6 gene in CAD ($p = 1.02003e - 15$, Fig 4c) and is associated with a decrease in gene expression. This data indicates that a longer TE might decrease the expression of RBM6.

3.3 Identifying the Affected Genes

The identified set of CAD-SNP-SV-eQTL associations revealed which genes are correlated with the established associations. Among the 321 associations, 109 unique genes were identified. To explore the relation of the affected genes to the effect sizes of SNP-SV, SNP-CAD, and SNP-eQTL associations, the dataset was additionally filtered to keep only those associations with all three p-values satisfying the threshold of $p < 5 \times 10^{-8} (0.05/10^6)$. This value stems from there being approximately 1 million independent common genetic variants in the human genome, thus indicating genome-wide significance [4]. The remaining 27 associations were used to construct a heatmap of effect sizes, outlined in Fig 3.

The highest magnitudes are observed for the positive effect sizes of SNPs rs6728861 and rs114407963, both associated with the SV chr2_203034349_D_19E0M1. This data is highlighted in the first column of Fig 3 in red. These SNPs are also moderately associated with CAD and mildly associated with the upregulation of genes CEP63 and CERS2.

A substantial downregulation of the gene RBM6 is exhibited by the SNP rs62060550, indicated by the dark blue color in the third column. This SNP also shows a moderate positive effect size in its association with SVs chr16_75469418_I_B91MF and chr16:75469157-75469915_JOIN. The SNP shows a very mild effect size for CAD.

The heatmap also shows a large majority of the genes being downregulated, as indicated by the prevalence of blue color in the third column. A large chunk of these associations also show a negative effect size in the first column, indicating associations of SNPs with shorter SVs.

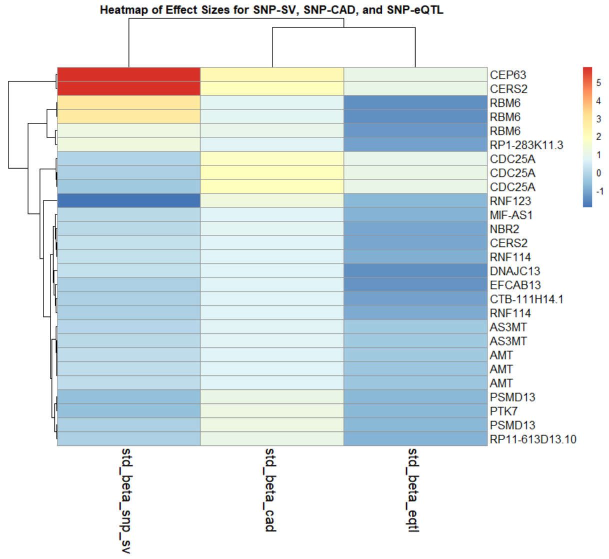


Figure 3: Heatmap of Effect Sizes for SNP-SV, SNP-CAD, and SNP-eQTL. The heatmap shows the standardized effect sizes of the associations with all p-values (SNP-SV, SNP-CAD, and SNP-eQTL) below 5e-08 in relation to genes. The blue-to-red color scale indicates the direction and magnitude of a given effect size.

3.4 GSEA Results

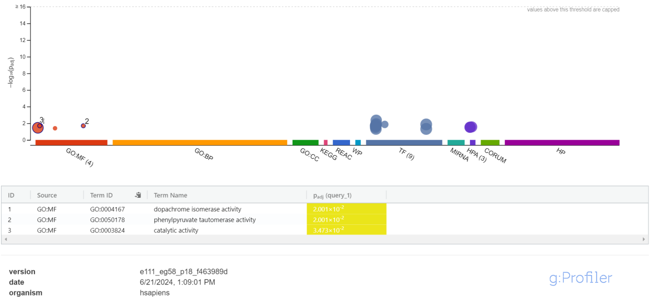
The 109 unique genes were used to perform gene set enrichment analysis to identify relevant biological pathways that might be affected. The results are shown in Fig 5.

Table 1: Top SNP-SV Associations

sv_chr	position	type	snp_id	p_snp_sv	p_cad	beta_cad	beta_snp_sv	score	NRA
9	22057960	TR	rs10811647	3.0e-133	1.0e-159	0.1471	-76.7748	282.77	164
1	169190783	TE	rs3737681	3.5e-286	1.6e-06	0.0275	-332.847	280.02	309
1	2994973	TE	rs946177	9.4e-282	2.5e-07	0.0438	-315.997	276.74	146
9	22057962	TE	rs10811650	3.2e-126	2.7e-160	0.1457	76.5354	276.14	166
18	23577153	TR	rs1624695	7.9e-277	1.2e-06	0.0293	-157.847	271.01	26
6	58155436	TR	rs1891819	2.2e-271	1.4e-06	0.0441	391.85	265.73	40
6	58155489	TE	rs1891819	2.4e-271	1.4e-06	0.0441	392.206	265.70	40
17	2153568	TR	rs9898819	1.7e-263	7.7e-12	0.0415	-77.7801	263.04	283
12	111858638	TE	rs6489836	2.8e-239	2.8e-09	0.0539	-723.202	236.91	143
20	59148390	TE	rs6026728	6.5e-235	9.4e-07	0.0453	335.718	229.45	112
20	35149142	TE	rs1415771	4.0e-217	9.9e-11	0.0328	616.417	216.05	180
13	112999855	TR	rs4907479	5.2e-208	4.0e-07	0.0339	-64.7433	202.77	49
20	52526106	TR	rs6091540	1.3e-207	8.4e-06	0.0248	-76.7924	200.71	1
4	119323404	TE	rs13108589	5.0e-204	6.8e-06	0.0272	-347.714	197.39	604
14	75032864	TE	rs8013780	5.4e-200	6.3e-09	0.0331	330.406	196.83	248
18	23577368	TE	rs1652349	3.0e-200	9.2e-07	0.0296	-156.524	194.69	26
2	63552088	TE	rs2422011	3.0e-200	6.2e-06	0.027	-326.309	193.99	0
1	3037749	TR	rs12034573	3.0e-199	8.2e-07	0.0361	-71.7238	193.60	183
13	112961877	TR	rs35872678	9.9e-196	5.2e-06	0.0284	64.3335	189.16	49
6	96569795	TE	rs11153071	2.1e-191	2.1e-06	0.0348	-321.086	185.43	43
3	69793492	TR	rs113374742	1.3e-189	6.7e-06	0.0304	-52.175	182.99	26
15	78874812	TE	rs62013185	3.0e-183	1.4e-10	0.052	331.582	181.83	283
2	164150135	TR	rs6731923	1.2e-182	6.6e-10	0.0455	54.2793	180.48	139
17	42106636	TE	rs35370188	5.4e-183	7.7e-08	0.0469	107.5	178.58	60
12	111858638	TE	rs7134638	1.3e-180	7.1e-09	0.0524	-725.992	177.95	222
20	59181784	TE	rs11908189	5.5e-183	1.0e-06	0.0441	117.253	177.38	112
10	103337648	TR	rs11191672	1.5e-181	5.2e-06	0.0232	311.487	175.30	1

sv_chr: Chromosome of the structural variant.
position: Starting position of the structural variant.
type: Type of structural variant (TR: Tandem Repeat, TE: Transposable Element).
snp_id: Identifier of the SNP.
p_snp_sv: Corrected p-value for SNP-SV association.
p_cad: Corrected p-value for SNP-CAD association.
beta_cad: Beta value for SNP-CAD association.
beta_snp_sv: Beta value for SNP-SV association.
score: Composite score from Fisher's method.
NRA: Number of Removed Associations from the genomic filtering step.

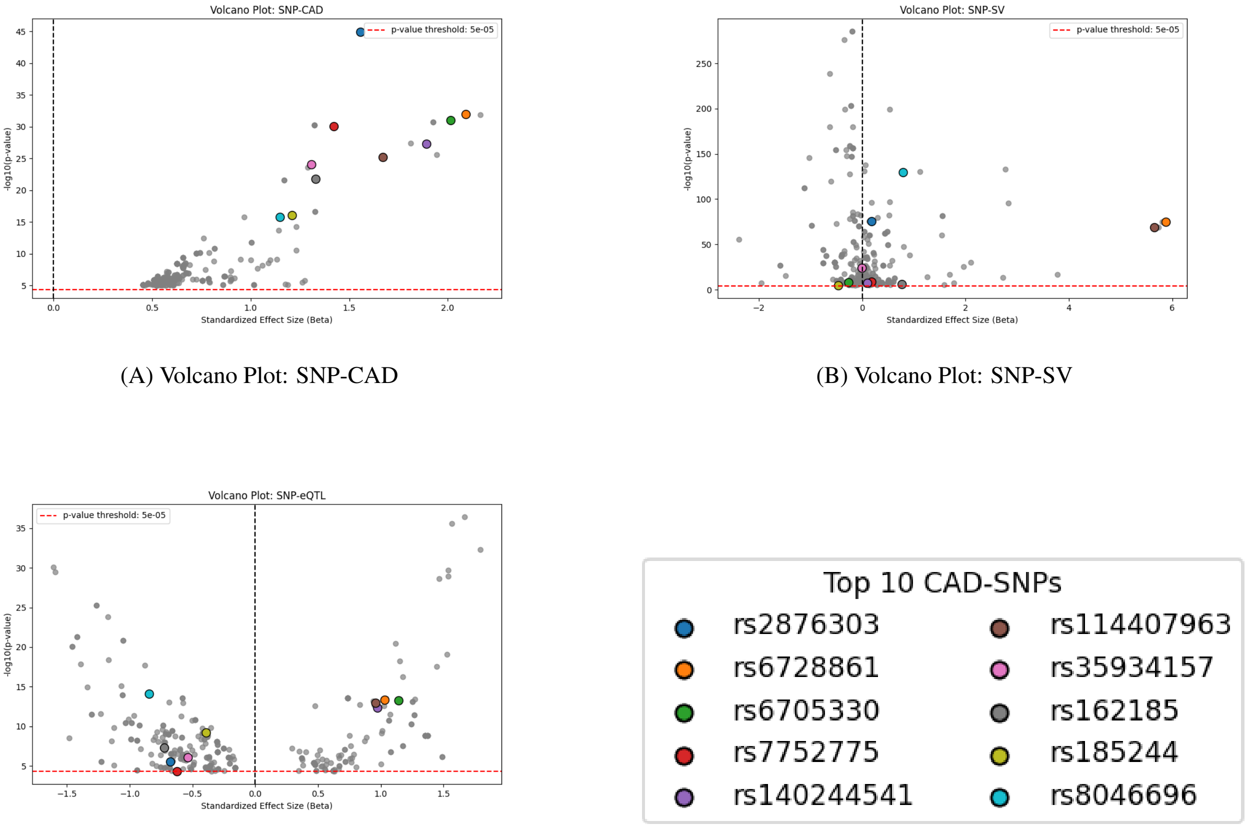
The results indicate GO:MF (Gene Ontology: Molecular Function) enrichment, specifically dopachrome isomerase activity (GO:0004167, $p = 2.001$e-2), phenylpyruvate tautomerase activity (GO:0050178, $p = 2.001$e-2) and catalytic activity (GO:0003824, $p = 3.473$e-2). Less notable enrichment can be seen in the Transcription Factors (TF) and Human Protein Atlas (HPA) categories.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0004167	dopachrome isomerase activity	2.001×10^{-2}
2	GO:MF	GO:0050178	phenylpyruvate tautomerase activity	2.001×10^{-2}
3	GO:MF	GO:0003824	catalytic activity	3.473×10^{-2}

version	e111_eg58_p18_f463989d
date	6/21/2024, 1:09:01 PM
organism	hsapiens

Figure 5: g:Profiler Gene Set Enrichment Analysis (GSEA) results. Enriched gene sets are shown across various categories, such as Gene Ontology Molecular Function (GO:MF), Biological Process (GO:BP), and KEGG pathways. Each dot represents a gene set with its corresponding $-log_{10}$(p-value). Significant gene sets are highlighted. The table below lists the top significant terms, their sources, term IDs, term names, and adjusted p-values.



(A) Volcano Plot: SNP-CAD

(B) Volcano Plot: SNP-SV

(C) Volcano Plot: SNP-eQTL

(D) Volcano Plot Legend

Figure 4: Volcano plots showing the relationship between the standardized effect size (x-axis) and p-values (y-axis) for the different associations: **(A)** The plot shows SNP-CAD associations, with the most significant ones highlighted by colors shown in (D). **(B)** The plot shows SNP-SV associations, highlighting the same SNPs as (A). **(C)** The plot shows SNP-eQTL associations, again highlighting the same SNPs as (A). **(D)** The legend indicating the colors used to display the most significant SNP-SV associations.

4 Discussion

4.1 Interpretation of Results

In the resulting data, TEs accounted for the majority of the CAD-SNP-SV associations (517/968). TEs were also responsible for the removal of a larger number of associations during filtering (23316) compared to TRs (14381). This may imply that TEs are more significant for CAD risk than TRs.

Interestingly, the data showed an equal frequency of positive and negative SNP-SV effect sizes. The same frequency was roughly reflected in the 27 most significant associations. This means there is no clear implication for how the length of a SV impacts its association with CAD.

Integrating eQTLs with the dataset revealed that 109 genes associated with CAD are being affected. SNPs rs6728861 and rs8046696 were shown to be highly significant across all three datasets. They indicated SVs chr2_203034349_D_19E0M1 (TE) and chr16_75395934_D_B8FMF (TE), which serve as examples of SVs possibly affecting CAD risk. The data implied that longer TEs might increase the expression of gene CEP63 and decrease the expression of RBM6. The heatmap revealed additional SNPs associated with the downregulation of RBM6 and also associated with longer SVs. This further supports the presumed effect of longer TEs on its expression. On the other hand, CEP63 and CERS2 stood out in the heatmap due to their correlation with SNPs that show very high effect sizes for the SV chr2_203034349_D_19E0M1 (TE). The same SNPs show significant effect size for CAD, implicating chr2_203034349_D_19E0M1 as a stand-out candidate SV, likely to be correlated with CAD. Furthermore, the heatmap showed a large chunk of associations linking shorter SVs with the downregulation of genes. However, it was established that there is no clear correlation between SV size and CAD risk. This might mean the downregulation of this group of genes isn't of major significance for CAD, but the full extent of the implications is uncertain.

A similar uncertainty is present for the results of GSEA, as it is difficult to draw conclusions about the implications for CAD from the highlighted enrichments.

4.2 Integration with Existing Literature

Despite their speculated importance [19], the available literature studying the relationship between SVs and CAD is very scarce. This is largely due to technological limitations and the explained complexities of SVs. Many studies fail to report SVs as a result, primarily focusing on SNPs instead [20][21]. We highlighted specific SVs, such as chr2_203034349_D_19E0M1 (TE) and chr16_75395934_D_B8FMF (TE) that might assist in closing the knowledge gap in this area.

There have been several studies linking SVs to specific genes associated with CAD [22]. One such gene is LDLR, which has been associated with specific TE deletions [23], but the same gene wasn't identified in our dataset. A study investigating the effect of the downregulation of CERS2 on myoclonic epilepsy identified a deletion on the first chromosome [24]. We identified the upregulation of the same gene being associated with a deletion on the second chromosome, chr2_203034349_D_19E0M1. This demonstrates the notion that SVs may impact various diseases through their regulatory function. The combination of changes in the expression of genes like CERS2 can have significant upstream effects on CAD pathogenesis, and SVs seem to significantly contribute to these changes.

4.3 Limitations

The results indicated that most associations achieved high composite scores due to their low SNP-SV p-values. In fact, for many of these associations, the remaining values had barely any contribution to the resulting score. Similarly, while standardizing the effect sizes allowed for direct comparison of the different datasets, it also significantly reduced the individual values, making their contribution to the score negligible across the board. This calls into question the assumption that the established associations are the most significant ones.

Another limitation is the handling of the SVs across the dataset. Two SVs were considered different unless their IDs matched exactly. This is largely due to their complexity and a lack of a standardized representation method. Even a difference of a few base pairs between the positions of two SV IDs would distinguish them as separate, despite this not being the case. This leads to some SNPs being overrepresented in the final dataset, seemingly linked with more SVs than expected. As a result, they might overshadow other relevant SNPs or skew the statistics acquired from the dataset.

4.4 Improvements and Future Work

The limitations caused by the oversight of the SNP-SV p-value's effect on the composite score can be addressed by adjusting their impact on the final score. One way of achieving this could be to introduce a weighted variant of the composite score, assigning weights to each of the terms. This would allow for balancing the impact of different terms in a preferred way. The initialization of these weights might prove challenging, however. Another approach could be using a different method for combining the p-values or to use separate terms for the individual p-values instead.

Another improvement can be made to the handling of SVs. When applying the genomic position filtering, a more complex matching condition can be implemented to try to circumvent different labels representing the same SV. This algorithm would have to take into account the rough position of the variant and check the similarity of the other association statistics, to conclude whether the same SV is being considered. Alternatively, a similar procedure can be applied to the initial set of associations. SVs having multiple distinct annotations in the dataset can be linked to a unique identifier, which would then be used in the genomic position filtering step as the match term.

4.5 Conclusion

In this report, we combined multiple datasets to identify SVs possibly associated with CAD risk. We ranked and filtered the associations based on a computed composite score to establish a list of those highly significant for CAD. The results revealed that TEs accounted for the majority of highly significant associations. Furthermore, our analysis indicated that the size of SVs did not have significant implications for CAD.

Integrating the data with eQTLs revealed 109 CAD-associated genes in high correlation with a SV. SNPs rs6728861 and rs8046696 were found to be associated with SVs chr2_203034349_D_19E0M1 and chr16_75395934_D_B8FMF respectively. They also showed significance across all three datasets. Genes CEP63, CERS2, and RBM6 were highly associated with SNPs that displayed relevant effect sizes in all datasets. There was no clear implication for the affected biological pathways.

Overall, the significant association of SVs with CAD-SNPs indicates a meaningful correlation with CAD. We provided an overview of such relevant SVs and the affected CAD genes, but more research in the area is required to reveal the full extent of the role of SVs in the development of CAD.

5 Responsible Research

As for the ethical considerations of this research, the data sources that were used all come from well-established sources and studies, which are presumed to have followed appropriate ethical guidelines. The data they provide is not identifiable for specific individuals, as it only contains large-scale statistical information. The results of this research follow suit, as they don't reveal any ethically compromising data. Furthermore, all of the computations were run locally, having minimal environmental impact. This means there are little to no ethical drawbacks to this study.

The data sources are publicly accessible, except for the SNP-SV dataset which is yet to be formally published. This means that the results are currently not reproducible, but will become so in the future, once this data is made available. The reproducibility of this research is supported by a detailed overview of the methodology and transparency regarding assumptions and decisions, outlined in this report. Additionally, several limitations have been addressed, acknowledging the shortcomings of the results. The full set of results made available in Appendix B further supports the reproducibility of the research.

References

[1] Zhifen Chen and Heribert Schunkert. Genetics of coronary artery disease in the post-gwas era. *Journal of Internal Medicine*, 290(5):980–992, 2021.

[2] Rachel Hajar. Risk factors for coronary artery disease: historical perspectives. *Heart views*, 18(3):109–114, 2017.

[3] Ruth McPherson and Anne Tybjaerg-Hansen. Genetics of coronary artery disease. *Circulation research*, 118(4):564–578, 2016.

[4] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

[5] Zhanye Zheng, Dandan Huang, Jianhua Wang, Ke Zhao, Yao Zhou, Zhenyang Guo, Sinan Zhai, Hang Xu, Hui Cui, Hongcheng Yao, et al. Qtlbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic acids research*, 48(D1):D983–D991, 2020.

[6] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[7] Colby Chiang, Alexandra J Scott, Joe R Davis, Emily K Tsang, Xin Li, Yungil Kim, Tarik Hadzic, Farhan N Damani, Liron Ganel, GTEx Consortium, et al. The impact of structural variation on human gene expression. *Nature genetics*, 49(5):692–699, 2017.

[8] Pim Van Der Harst and Niek Verweij. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, 122(3):433–443, 2018.

[9] GTEx Consortium Lead analysts: Aguet François 1 Brown Andrew A. 2 3 4 Castel Stephane E. 5 6 Davis Joe R. 7 8 He Yuan 9 Jo Brian 10 Mohammadi Pejman 5 6 Park YoSon 11 Parsana Princy 12 Segrè Ayellet V. 1 Strober Benjamin J. 9 Zappala Zachary 7 8, NIH program management: Addington Anjene 15 Guan Ping 16 Koester Susan 15 Little A. Roger 17 Lockhart Nicole C. 18 Moore Helen M. 16 Rao Abhi 16 Struewing Jeffery P. 19 Volpi Simona 19, Pathology: Sobin Leslie 30 Barcus Mary E. 30 Branton Philip A. 16, NIH Common Fund Nierras Concepcion R. 137, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[10] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[11] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[12] Wiesje M Van Der Flier and Philip Scheltens. Amsterdam dementia cohort: performing research to optimize care. *Journal of Alzheimer's Disease*, 62(3):1091–1111, 2018.

[13] Henne Holstege, Nina Beker, Tjitske Dijkstra, Karlijn Pieterse, Elizabeth Wemmenhove, Kimja Schouten, Linette Thiessens, Debbie Horsten, Sterre Rechtuijt, Sietske Sikkes, et al. The 100-plus study of cognitively healthy centenarians: rationale, design and cohort description. *European Journal of Epidemiology*, 33:1229–1249, 2018.

[14] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[15] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[16] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[17] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[18] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[19] Ali J Marian and John Belmont. Strategic approaches to unraveling genetic causes of cardiovascular diseases. *Circulation research*, 108(10):1252–1269, 2011.

[20] Ishani Mhatre, Habiba Abdelhalim, William Degroat, Shreya Ashok, Bruce T Liang, and Zeeshan Ahmed. Functional mutation, splice, distribution, and divergence analysis of impactful genes associated with heart failure and other cardiovascular diseases. *Scientific reports*, 13(1):16769, 2023.

[21] Nabila Bouatia-Naji, Keren J Carss, Javier Armisen, Tom R Webb, Stephen E Hamby, Diluka Premawardhana, Abtehale Al-Hussaini, Alice Wood Bsc, Quanli Wang, Sri VV Deevi, et al. Spontaneous coronary artery dissection. *Circulation: Genomic and Precision Medicine*, 13(6), 2020.

[22] Bram P Prins, Vasiliki Lagou, Folkert W Asselbergs, Harold Snieder, and Jingyuan Fu. Genetics of coronary artery disease: genome-wide association studies and beyond. *Atherosclerosis*, 225(1):1–10, 2012.

[23] Maria S Nazarenko, Aleksei A Sleptcov, Aleksei A Zarubin, Ramil R Salakhov, Alexander I Shevchenko, Narek A Tmoyan, Eugeny A Elisaphenko, Ekaterina S Zubkova, Nina V Zheltysheva, Marat V Ezhov, et al. Calling and phasing of single-nucleotide and structural variants of the ldlr gene using oxford nanopore minion. *International Journal of Molecular Sciences*, 24(5):4471, 2023.

[24] Mai-Britt Mosbech, Anne SB Olsen, Ditte Neess, Oshrit Ben-David, Laura L Klitten, Jan Larsen, Anne Sabers, John Vissing, Jørgen E Nielsen, Lis Hasholt, et al. Reduced ceramide synthase 2 activity causes progressive myoclonic epilepsy. *Annals of clinical and translational neurology*, 1(2):88–98, 2014.

A Genomic position filtering algorithm pseudocode

Algorithm 1 Genomic Filtering Based on Position

```
for $i \leftarrow 1$ to $len(combined\_associations)$ do
    $current\_row \leftarrow combined\_associations[i]$
    $current\_pos \leftarrow current\_row['pos']$
    $current\_sv\_id \leftarrow current\_row['sv\_id']$
    $window\_start \leftarrow current\_pos - 250000$
    $window\_end \leftarrow current\_pos + 250000$
    for $j \leftarrow i + 1$ to $len(combined\_associations)$ do
        $comparison\_row \leftarrow combined\_associations[j]$
        if $comparison\_row['sv\_id'] == current\_sv\_id$ and $window\_start \leq comparison\_row['pos'] \leq window\_end$ then
            Remove $comparison\_row$ from $combined\_associations$
        end if
    end for
end for
```

B Supplemental information

You can download the Excel sheet containing the full results here.