KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCES

Novel method for annotating gene targets to GWAS loci

Promoter:
Prof. Peter Claes
Department of Human Genetics
Department of Electrical Engineering, ESAT/PSI

Co-promoter:
Prof. Isabelle Cleynen
Department of Human Genetics

Dissertation presented in
fulfillment of the requirements
for the degree of Master of Science:
Bioinformatics

Alita D'HOORE

June 2025

Foreword

I would like to thank my promotor, Prof. Peter Claes, for giving me the opportunity to work on this thesis. I would also like to express my sincere gratitude to my supervisor, Seppe Goovaerts, who guided me throughout the entire process. Thank you for always being available to answer my questions and for inspiring me with your knowledge and creativity.

Lastly, I would like to thank my family, particularly my mother, whose support throughout my educational career and constant belief in me have been invaluable.

Abstract

Genetic variation among individuals partially explains phenotypic differences between them. Most of this variation is identified in the form of Single Nucleotide Polymorphisms (SNPs), which are statistically tested for phenotype associations through Genome-Wide Association Studies (GWAS). Complex traits and diseases are typically associated with hundreds of SNPs, the majority of which lie in non-coding regions of the genome. Rather than altering the structure or function of proteins directly, these non-coding SNPs often influence regulatory elements that affect transcription and gene expression. Importantly, such regulatory SNPs are not necessarily located near the genes they influence; they may reside in distal enhancer regions that interact with promoter regions through the three-dimensional architecture of the genome.

GWAS identify phenotype-associated loci, and the next critical step is to determine the target genes of these loci in order to uncover and understand the underlying biological mechanisms. The most widely used gene-annotation strategy is the distance-based approach, which assigns each SNP to its nearest gene in linear genomic space. However, this method is limited by the three-dimensional structure of the genome, which allows distant regions to come into close spatial proximity and interact. These interactions are often organized within Topologically Associating Domains (TADs), which define regions of the genome that frequently interact.

In this thesis, a novel gene-annotation strategy was developed to address the limitations of proximity-based methods. Expectation-Maximization for the Annotation of Genes using GO enrichment (E-MAGO) is a method that limits the SNP search space using TAD boundaries and iteratively scores candidate genes based on Gene Ontology (GO) enrichment, thereby prioritizing genes based on biological context.

E-MAGO was first validated against the distance-based method using disease-associated genes from GeneCards as ground truth. This comparison showed that E-MAGO outperformed the distance-based approach, particularly in gene-dense regions where proximity-based methods tend to misassign SNPs to nearby but biologically irrelevant genes. Next, genes linked to lead SNPs through expression quantitative trait loci (eQTL) overlap were used as an alternative ground truth. In this case, the distance-based method outperformed E-MAGO. However, this result likely reflects limitations of the validation method rather than E-MAGO itself, as it relied on simple overlap with eQTLs, which may capture effects from variants in linkage disequilibrium rather than direct regulatory relationships, reducing the specificity of the ground truth.

Overall, this thesis demonstrates that E-MAGO offers a biologically informed alternative to traditional annotation methods by integrating 3D genome architecture and functional gene context. While challenges remain in validation, E-MAGO shows promise for improving the identification of target genes at GWAS loci and for enhancing our understanding of the biological mechanisms underlying complex traits and diseases.

List of Abbreviations and Symbols Used

BC	breast cancer
CTCF	CCCTC-binding factor protein
DAG	directed acyclic graph
E-MAGO	expectation-maximization for annotation of genes using GO enrichment
EM	expectation-maximization
eQTL	expression quantitative trait locus
GO	gene ontology
GOEA	gene ontology enrichment analysis
GREAT	genomic regions enrichment of annotations tool
GWAS	genome-wide association studies
Hi-C	high-throughput chromosome conformation capture
IBD	inflammatory bowel disease
kb	kilobases
LD	linkage disequilibrium
MAF	minor allele frequency
Mb	megabases
QC	quality control
SNP	single nucleotide polymorphism
T2D	type 2 diabetes
TAD	topologically associating domain
TSS	transcription start site

List of Tables

List of Figures

Table of Contents

1 Literary Review

1.1 Genome Architecture

At the core of human biology lies a complex genome, whose structure and subtle variations form the basis of various human traits and diseases. This genome is built out of nucleotides which are attached one after another to form polymers called DNA strands. The human genome consists of over 3 billion nucleotides inherited from each parent. (Fraser et al., 2015) Nucleotides can have different bases, namely adenine, cytosine, guanine, or thymine, which are organized in a specific way to create genes and intergenic regions that separate them. (Nygaard & Saxild, 2009) The sequence of a gene is subdivided into groups of three nucleotides which are called codons. When DNA is transcribed into mRNA these codons will translate into amino acids. The combination of these amino acids forms a protein. This mechanism describes the central dogma of molecular biology (Figure 1) which states that the genetic code moves unidirectional in that way. (Ostrander, 2025) While genes carry the instructions for building proteins, intergenic regions play a crucial regulatory role by housing elements such as promoters, enhancers, and silencers**,** which control when and where genes are expressed. In eukaryotic organisms like humans, genes are composed of exons, which code for proteins, and introns, which are non-coding sequences. During transcription, the entire gene—including both exons and introns—is transcribed into pre-mRNA. The introns are then removed through splicing, as they do not contribute to the final protein product. Although introns do not code for proteins, they serve several important regulatory functions. For example, they enable alternative splicing, allowing a single gene to produce multiple protein variants; they can enhance gene expression through intron-mediated mechanisms; and they may act as buffers, protecting coding regions from potentially harmful mutations. (Jo & Choi, 2015)

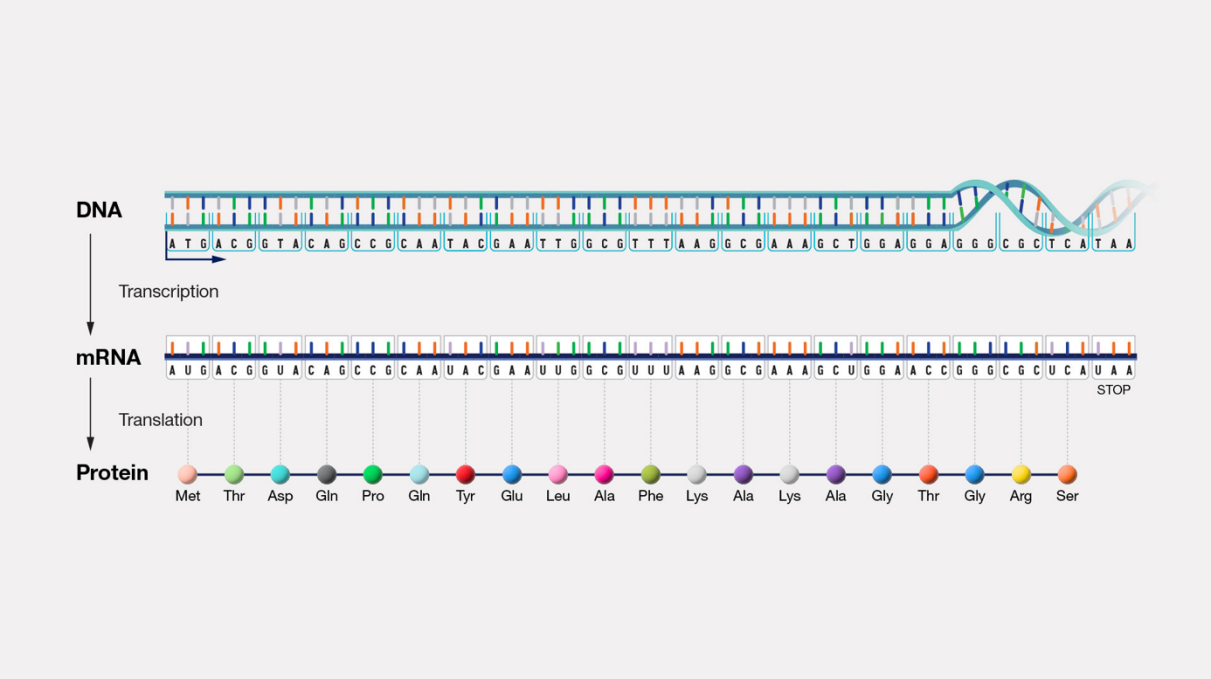


Figure 1: Visual representation of the central dogma of molecular biology. *(Ostrander, 2025)*

Facilitated by histones, the DNA is tightly folded into chromatin. Chromatin structure governs the accessibility of DNA sequences by modulating their spatial organization, influencing whether

regulatory elements can interact or remain inaccessible. These internal chromatin interactions are essential for defining active regions involved in DNA-dependent processes such as replication and transcription. Further packaging will make the chromatin form even more dense structures that are known as chromosomes. Chromosomes consist of three types of chromatins: centromeric chromatin, heterochromatin and euchromatin. The centromeric chromatin can be found at the centromere located at the center of a chromosome. It is crucial for correct chromosome segregation during mitosis. The heterochromatin is defined as relatively inaccessible chromatin due to compact folding. As a result, heterochromatin will be transcriptionally inactive or in other words: gene silencing. There are two types of heterochromatins: facultative which changes during different developmental stages and constitutive which remains constant. The third form of chromatin, euchromatin, is known to be very loose and open. It consists of active genes and therefore high accessibility for transcription. The promotor regions and regulatory elements essential for transcription near these active genes will also consist of euchromatin. (Morrison & Thakur, 2021)

Promotors are regions located close to (< 50 kilobases) the transcription start site (TSS) where they facilitate the initialization of transcription. By binding general transcription factors (GTFs) and recruiting the RNA polymerase to the promotor sequence, the pre-initiation complex (PIC) is formed. This complex is essential for initializing transcription. Often, additional signals are needed to fully activate transcription, which are called enhancers. Enhancers are sequences that can improve transcription through interaction with their target promoter. Independent of their distance or orientation (upstream or downstream), enhancers will bind regulatory proteins known as transcriptional activators, which in turn recruit co-activators such as the Mediator complex. These co-activators are brought into proximity to the promoter region through the 3D folding of the DNA, where they can interact with the PIC to activate transcription. (Haberle & Stark, 2018) Enhancers can be far from their target promoters in linear genomic distance but are brought into spatial proximity via DNA looping and higher-order chromatin organization. This 3D architecture allows enhancers to modulate gene expression with precision, bypassing linear constraints which is demonstrated in Figure 2. (Yang & Hansen, 2024) In addition, transcription factors can influence 3D genome organization by recruiting chromatin modifiers that alter DNA accessibility and chromatin structure. These changes can affect enhancer–promoter interactions and contribute to the formation of regulatory loops that enable precise gene expression. (Kim & Shendure, 2019)

Chromatin is organized into Topologically Associating Domains (TADs)—regions of the genome that interact more frequently within themselves than with neighboring regions. Within a TAD, functional elements such as genes and regulatory regions primarily interact with each other, playing a key role in establishing 3D enhancer–promoter interactions. This reflects a broader organizational principle in which the genome is partitioned into distinct compartments, where regulatory elements within the same TAD are more likely to interact and often correspond to co-regulated genes. (McArthur & Capra, 2021; Szabo et al., 2019) TADs are generally stable across cell types; however, TAD boundaries differ from the domains themselves in that they are more evolutionarily conserved and enriched in critical genomic features, such as housekeeping genes and TSSs**.** Disruption of these boundaries can lead to misregulation of gene expression, contributing to disease—for example, acute myeloid leukemia can result from pathogenic enhancer–gene mispairing. Overall, TADs are important genomic structures to interpret biologically meaningful interactions of functional elements like enhancer-promotor

complexes by confining them within structural domains. (McArthur & Capra, 2021)

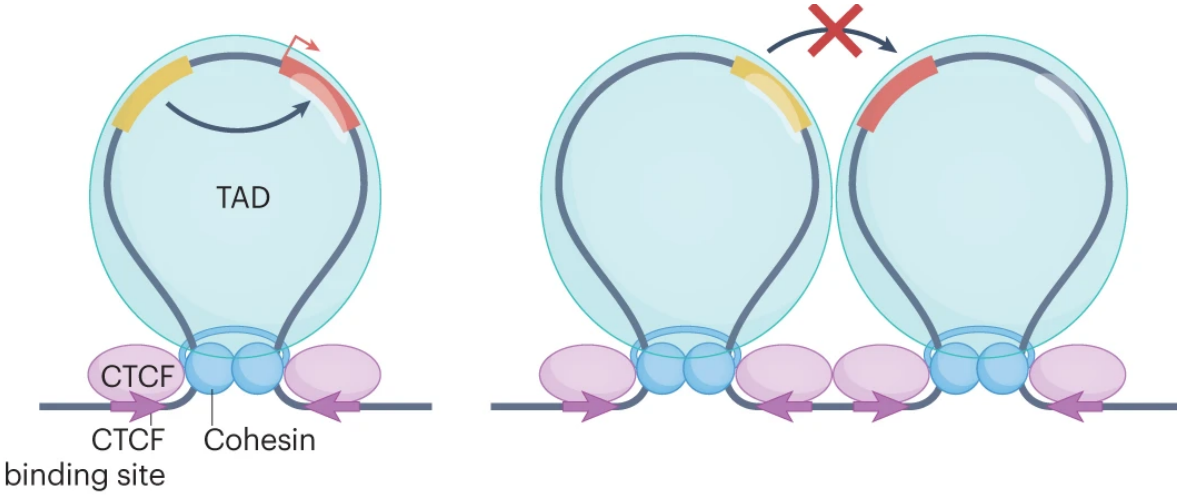


Figure 2: Mechanism of interactions within a TAD which is formed through CTCF and cohesin interaction. *(Yang & Hansen, 2024) The left figure demonstrates how an enhancer interacts with a promotor region inside a TAD and the right figure shows that such interactions between TADs are not possible.*

The formation of TADs is often explained by the loop extrusion model. In this model, the structural maintenance of chromosomes cohesin complex (SMC), a ring-shaped protein structure, plays a central role in shaping DNA loops. The cohesin complex attaches to the DNA and extrudes a loop by progressively pulling DNA through its ring. Loop boundaries are typically defined by the CCCTC-binding factor protein (CTCF); when cohesin encounters such CTCFs, extrusion halts, stabilizing the loop as seen in Figure 2. CTCF binding sites are often found near enhancers or promoters, which exhibit strong interactions following loop formation. However, not all loops require CTCF—some loops form independently, either when cohesin dissociates from the DNA or through strong enhancer-promoter interactions alone. (Szabo et al., 2019)

The 3D organization of these domains has been revealed through High-throughput Chromosome Conformation Capture (Hi-C), a high-throughput adaptation of chromosome conformation capture techniques. (Han et al., 2018) TADs are visualized using Hi-C interaction maps**,** where contact frequencies between genomic regions are color-coded—typically, darker colors indicate stronger interactions. A corner peak appears as a vertex of one of the squares on these maps (Figure 3). Such peaks often mark the end of a chromatin loop and highlight a point of strong interaction such as an enhancer-promotor interaction. (Krietenstein et al., 2020)

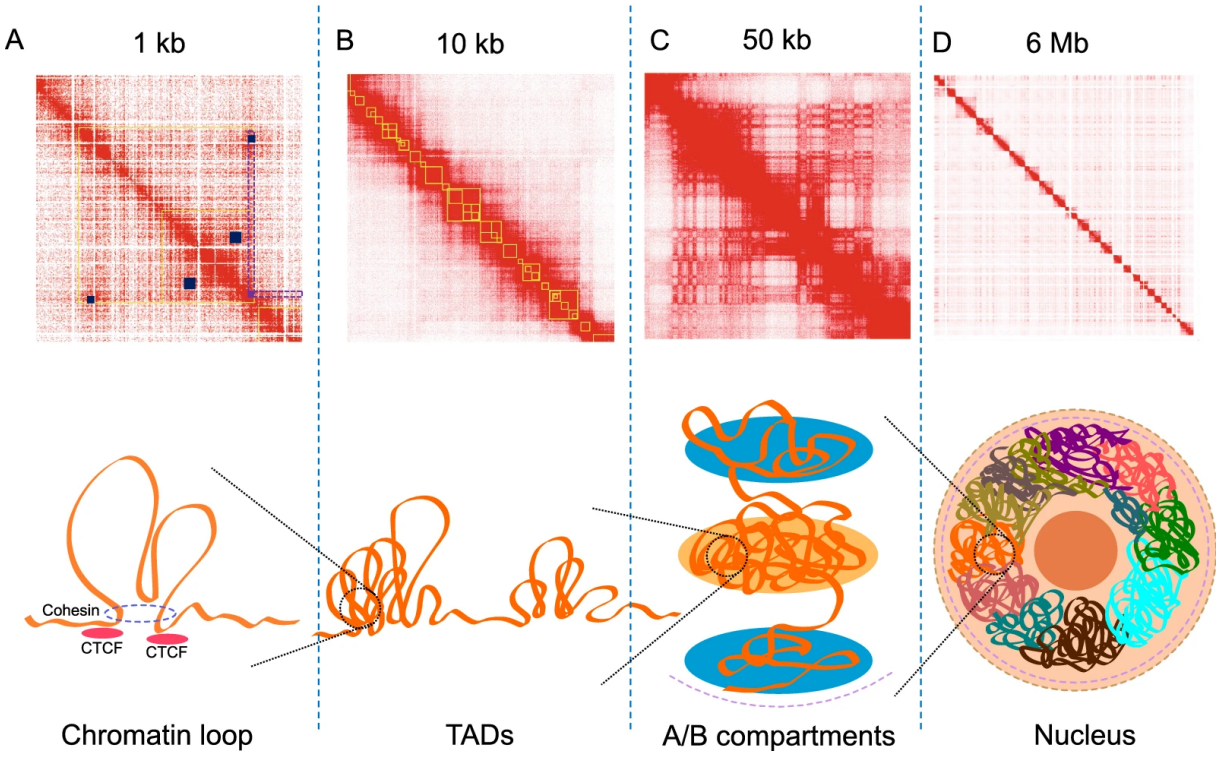


Figure 3: Different layers of the 3D genomic structure. *(N. Liu et al., 2021) The top row shows Hi-C interaction maps on different scales progressively zooming out on the 3D form. The bottom row shows the actual corresponding 3D architectures.*

1.2 Genetic variation: SNPs

Traits and diseases are partially encoded in a person's genetic code. Most of the time, the genetic variation amongst individuals partially explains the phenotypic differences between individuals. This genetic makeup contains DNA sequences of which only 0.1% uniquely identify individuals through genetic variations. The most common type of variation is a Single Nucleotide Polymorphism (SNP). SNPs are single base-pair substitutions that are spread throughout the genome with a rate of 1 appearing every 1000 nucleotides. (Komar, 2009) SNPs can occur in coding and non-coding parts of the genome. A SNP located within a coding region can directly impact a gene by altering the structure and function of the resulting protein. In some cases, the change in nucleotide does not affect the amino acid sequence due to the redundancy of the genetic code—this is known as a synonymous mutation. When the SNP does lead to an amino acid change, it is referred to as a non-synonymous mutation, which can be further categorized as a missense mutation (resulting in a different amino acid) or a nonsense mutation (introducing a premature stop codon). SNPs occurring in non-coding or intergenic regions can influence a variety of regulatory elements, such as promoters, enhancers, or splice sites. Although they do not alter protein sequence directly, these changes can significantly impact gene expression, RNA processing, or transcriptional activity. (Komar, 2009; Mocellin, 2007) Non-coding and intergenic SNPs are the most abundant, and although their functions remain less well understood than those of coding variants, their presence suggests that SNPs in general play an important role in regulating transcription and gene expression. (Zou et al., 2020) SNPs can serve as biological markers for associating genetic variations with specific traits and diseases. They are abundant throughout the genome, and the minor allele is typically found in more than 1% of the population. This frequency is referred to as the Minor Allele Frequency (MAF)**,** where a MAF greater than 5% indicates a common variant, and an MAF below 1% classifies a rare variant. Furthermore, SNPs that confer a selective

advantage to individuals tend to become genetically stabilized within populations over the course of evolution. (Barreiro et al., 2008)

1.3 Complex versus Mendelian inheritance

Traits and diseases can broadly be categorized in monogenic and polygenic. Monogenic traits are expressed through one or a few genes. Most monogenic traits can also be named mendelian. This implies that the trait is 100% determined by its genotype which follows from Mendel's laws of inheritance. According to Mendel's law of segregation, a trait has two alleles which will be separated in gametes where each gamete will have one allele from each gene and the alleles merge randomly in the progeny. (Strome et al., 2024) Also, Mendel's law of independent assortment states that alleles at different loci on the chromosome will segregate independently from each other into gametes. Mendel's laws describe the inheritance patterns of single genes with two alleles, typically associated with monogenic traits. These traits exhibit clear, discrete phenotypes, with alleles classified as either dominant or recessive. (Miko, 2008) When a parent carries two alleles, the child has a 50% chance of carrying either one of the alleles. If the child inherits a dominant allele, then the dominant trait will appear. Only when the child inherits two recessive alleles, one from each parent, then the recessive trait will appear. (Figure 4) In some cases, monogenic traits do not present with clear-cut, discrete phenotypes due to the influence of variable expressivity and incomplete penetrance**.** Variable expressivity refers to situations where individuals with the same genetic variant (e.g., SNP) exhibit different degrees of the phenotype. On the other hand, incomplete penetrance occurs when some individuals with the variant show no phenotype at all. (Kingdom & Wright, 2022)

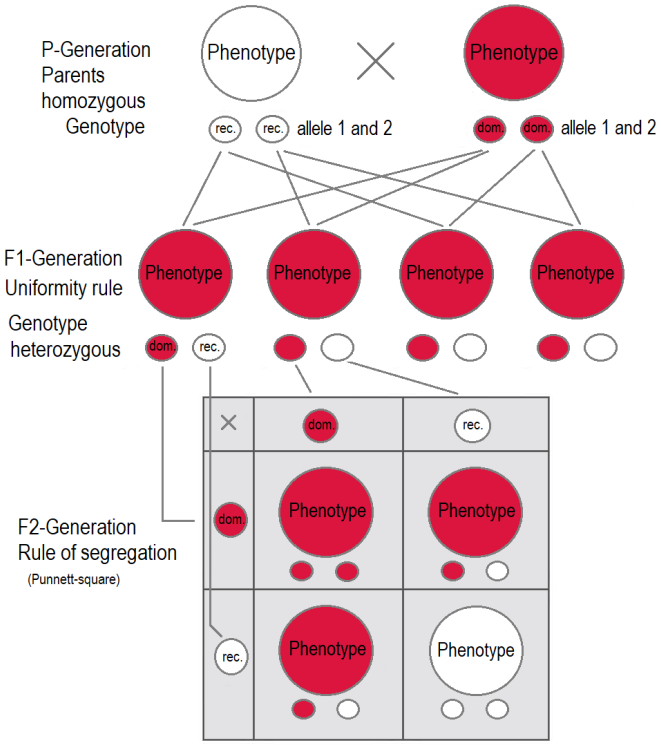


Figure 4: Schematic overview of the mendelian inheritance pattern. *(*Mendelian Inheritance - Wikipedia*, n.d.) On top, the different crossings show the different allele combinations possible for the F1-generation. At the bottom, F2-generation is visualized with a Punnett square.*

In contrast, complex traits are polygenic traits with every genetic variant contributing a small effect to the overall phenotype, along with additional mechanisms such as environmental influences (e.g., diet, lifestyle, or exposure to toxins), gene–environment interactions (where the effect of a gene depends on specific environmental conditions), gene–gene interactions (where the combined effect of genes differs from the sum of their individual effects), and stochastic processes—random biological fluctuations during development or gene expression that can cause variability even among genetically identical individuals. (Manolio et al., 2009; Valdar et al., 2006) This complex architecture complicates the prediction of inheritance patterns, as it does not conform to the simple Mendelian model. Instead of discrete categories, complex traits display a continuous spectrum of phenotypic variation. Furthermore, the involvement of many genes increases the likelihood that some will be physically linked on the same chromosome. This means that alleles close to each other on the same chromosome are not inherited independently but are rather strongly associated which is known as Linkage Disequilibrium (LD). (Kockum et al., 2023) This linkage violates Mendel's law of independent assortment, further adding complexity to the inheritance patterns of complex traits where alleles in LD are inherited together. (Crouch & Bodmer, 2020) In addition to their complex inheritance patterns, another distinguishing feature of these traits lies in the genomic locations of their associated genetic variants. While Mendelian traits are often caused by rare, high-impact mutations within coding regions of a single gene, most SNPs associated with complex traits are located in non-coding regions of the genome. (Boyle et al., 2017)

Well-studied traits that have a polygenic architecture are coronary artery disease (CAD) and Crohn's disease (CD). For CAD, there are 48 lead SNPs associated to the disease. The study investigated 10 novel loci where all 10 of these SNPs are shown to be residing in intronic or non-coding regions. (Nikpay et al., 2015) For CD, there are 119 lead SNPs where the conclusion can be made that most of the SNPs are in non-coding regions. (Z. Liu et al., 2023)

In contrast, monogenic traits that have only one or a few lead SNPs affecting one or a few genes are Cystic Fibrosis (CF) and Sickle Cell Disease (SCD). CF is associated with SNPs located in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTCR) gene. (Ong & Ramsey, 2023) SCD, on the other hand, is caused by a single SNP rs334 which is located inside the β-globin gene. (Sundd et al., 2019) Both illustrating that for monogenic traits the variations do lie within coding regions of a gene.

1.4 Genome wide association studies

SNPs are identified using a variety of genotyping techniques. The choice of method depends on several factors, including the number of SNPs to be mapped, the size of the cohort, computational capacity, and cost-effectiveness. For example, CF is a monogenic trait where primarily SNPs in the CFTR gene need to be analyzed. This represents a targeted study with a relatively small cohort size, where high computational power is not essential. In such targeted studies, PCR-based methods are often used, as they allow for the detection of specific alleles within a preselected DNA region, enabling the association of particular variants with the trait or disease. In contrast, for polygenic traits like CD, large cohorts and genome-wide SNP genotyping are required. In these cases, sequencing methods provide the most detailed representation of the genome and the variation across loci between individuals. Although sequencing offers the advantage of identifying novel SNPs, it also generates large datasets that require extensive manual quality checking, which can be time-consuming. An alternative approach is the use of genotyping arrays, which allow for the simultaneous genotyping of large portions of the

genome. The most widely used platform is the microarray, which consists of microscopic spots containing DNA probes complementary to known SNPs of interest. When a fragment of the targeted DNA hybridizes to a probe, single base extension occurs, incorporating a fluorescently labeled nucleotide complementary to the SNP. The resulting fluorescent signal is detected, revealing which allele is present at the given locus. Finally, several quality control and imputation steps are applied to ensure the reliability and accuracy of the genotyped data. (Kockum et al., 2023)

After SNP genotyping, genotypes can be linked to phenotypes through Genome-Wide Association Studies (GWAS)**.** This approach uses a regression model to test for statistical associations between genetic variants and traits or diseases with every genetic variant tested separately. Depending on whether the trait or disease is continuous or binary, a linear or logistic regression model is used, respectively. The phenotype is the dependent variable, and the SNP is portrayed as the independent variable which has a coefficient that explains the amount of impact the SNP has on the phenotype. The statistical power of the model is enhanced by incorporating additional terms. (Dehghan, 2018) First, covariates such as sex, age and ancestry are considered to avoid confounding as allele frequencies could differ among different groups of these covariates. Second, a random effect term can be added to correct for relatedness between individuals. Finally, because nearby SNPs are often in LD and not statistically independent, another term may be included to adjust for this polygenic background, capturing the influence of multiple small-effect variants across the genome. Additionally, a genetic inheritance model must be selected to describe how a single SNP influences the phenotype. A GWAS most often assumes an additive model as the true inheritance pattern is typically unknown. The additive model effectively captures scenarios where the risk increases linearly with each additional copy of the risk allele. (Dehghan, 2018; Uffelmann et al., 2021)

As every SNP is regressed and tested for association individually, it is crucial to control for false discovery to correct p-values that look significant by chance. Conducting a GWAS for 1 million SNPs with a standard p-value smaller than 0.05 would mean finding 50 000 false discoveries by chance. The most used multiple testing correction in GWAS is the Bonferroni correction. It divides the significance threshold by the number of tests. So, for a threshold of 0.05 and 1 million SNPs tested, this would result in a corrected significant threshold of $5*10^{-8}$. This strict threshold of $5*10^{-8}$ has become a standard for GWAS involving common variants as it effectively corrects for false positives, and it makes it easier to compare different studies on different scales. (Chen et al., 2021) Also, previous studies have shown that the human genome counts on average 1 million independent SNPs. (Uffelmann et al., 2021) The Bonferroni correction is simple and assures low false positives rates, but because of its strictness it may not detect certain true positive SNPs with small effects on the trait.

GWAS have evolved quickly in the past 20 years. From the first GWAS being conducted in 2005 associating age-related macular degeneration to 103 611 SNPs (Klein et al., 2005) to now over 5 000 traits and diseases investigated through GWASs that have been conducted with the possibility to investigate over 10 million SNPs. This is the result of major advances in sequencing technologies and SNP genotyping in combination with growing reference panels like HapMap that make it possible to conduct GWAS for millions of SNPs to associate to complex traits and diseases. (Loos, 2020)

It is important to understand that GWAS cannot directly identify which specific SNPs cause a trait or disease. Instead, the SNPs detected are often in LD with the true causal variant. The region containing

both the causal variant and its correlated SNPs is referred to as a trait-associated locus (for traits) or a genomic risk locus (for diseases). GWAS signals typically highlight these loci, even if the actual causal variant was not directly genotyped. With the increasing size of GWAS cohorts, researchers can now detect SNPs with much smaller effect sizes and uncover genomic loci that may have been missed in smaller studies. (Visscher et al., 2017) For example, GWAS have recently uncovered genetic loci associated with altered brain activation patterns during cognitive tasks in schizophrenia, revealing neurobiological mechanisms not detected before. In a study using functional MRI during an auditory oddball task, GWAS identified novel SNPs significantly associated with reduced activation in the right supramarginal gyrus—a region implicated in attention and cognitive processing. One of the strongest associations was found for the SNP rs73200372, located in a non-coding region on chromosome 12, highlighting the importance of regulatory variation. (Nakahara et al., 2023)

The discovery of new loci associated with complex traits and diseases further expands the understanding of their polygenic architecture. Most SNPs are common variants with small effect sizes. For such complex diseases, a Polygenic Risk Score (PRS) can be calculated to estimate an individual's susceptibility to the disease. This score is determined by summing the number of risk alleles an individual carries, each weighted by its effect size. Most individuals in the population carry some of these SNPs, which typically confer only a low risk for the disease. However, the more risk alleles a person carries—and the larger their effect sizes—the greater the overall risk, particularly when combined with the right environmental factors. (Lewis & Vassos, 2020)

Although GWAS has led to a substantial increase in the number of genomic loci associated with complex traits and diseases, these loci account for only a small fraction of the total genetic heritability. It is certain that in the future an increase in sample size will further increase the number of loci associated to the trait or disease. (Visscher et al., 2017) In contrast, family-based studies, where familial clustering is considered, typically explain a much larger proportion of heritability. This discrepancy is referred to as missing heritability. For example, while genetic heritability of height is estimated to be around 80% based on family studies, current GWAS findings explain only 27.4%. (Génin, 2020) One reason for this gap is that a GWAS often filters for statistically significant SNPs, excluding many variants with small effect sizes. Including all SNPs, regardless of their individual significance, increases the proportion of heritability explained. This is further supported by the continued growth in GWAS sample sizes, which has enabled the discovery of additional loci that contribute to genetic heritability. A broader perspective on this issue is the shift from viewing complex traits as polygenic—influenced by many genes—to omnigenic, a model suggesting that nearly all genes expressed in relevant cell types may contribute indirectly to a trait or disease. (Boyle et al., 2017) Additionally, rare variants with strong effects may not be captured by SNP genotyping arrays, contributing further to the unexplained heritability. Finally, interactions between genes (epistasis) and between genes and environmental factors (gene–environment interactions) are typically not accounted for in standard GWAS models, but they likely play a significant role in shaping complex phenotypes. (Génin, 2020)

1.5 Gene annotation strategies

Identifying genomic loci associated with specific phenotypes is only the first step toward understanding the biological mechanisms underlying traits and diseases. Post-GWAS analyses aim to

investigate the functional relevance of each associated locus, including how these loci might influence proteins that contribute to the phenotype. However, since the majority of SNPs linked to complex traits lie in non-coding regions, they are more likely to affect gene regulation rather than directly altering protein structure. This makes it challenging to determine the functional impact of a given locus—particularly to answer the key question: *"Which gene is affected?"* Therefore, identifying the specific genes influenced by trait-associated loci is essential for uncovering the biological pathways involved. (Gallagher & Chen-Plotkin, 2018) These insights not only deepen our understanding of disease mechanisms but also helps the development of more targeted therapies.

1.5.1 Distance-based methods

The most popular way for annotating genes to loci is by simply taking the one in closest linear proximity. However, this strategy does not consider any functional or biological relation between the gene and locus. (Gazal et al., 2022) The downfall for these types of methods is the presence of distal regulatory elements. The locus close to a gene could be an enhancer for a gene much further away. The 3D-structure of the DNA makes it possible to bring such enhancer and gene closer to each other by folding their regions closer to each other. More than half of enhancers are at least 10 kilobases (kb) and up to 1.9 Megabases (Mb) away from their target genes. High distances are observed in extreme situations like gene deserts which are known to be enhancer dense regions that act on genes a few hundred kb away. (Brodie et al., 2016; Spitz, 2016) Annotating loci using distance results in an approximate precision and recall of 0.34 as shown in Figure 5. This means that only 34% of the genes predicted as closest to the SNPs are actually correct (precision) and correctly identify the true target genes (recall). However, this strategy remains widely used in many studies due to its simplicity and computational efficiency. Additionally, the continued reliance on it is often driven by the lack of structural or functional data needed to identify long-range regulatory interactions—particularly in challenging contexts such as brain tissue or tissue from developmental stages, where data are difficult to obtain, or in understudied phenotypes. (Brodie et al., 2016)

Several tools that use a distance-based strategy are Genomic Regions Enrichment of Annotations Tool (GREAT) and FUnctional Mapping and Annotation (FUMA). GREAT functionally annotates genomic regions by associating Gene Ontology (GO) terms with regulatory regions of genes. In a first step these genomic regions are annotated to genes by assigning regulatory regions to genes and finding the closest overlapping ones with the input genomic regions. This assignment is solely based on distance where the regulatory regions define how far away the gene can be from the input region. The standard setting is 5 kb upstream and 1 kb downstream from the transcription start site. In case no overlap is found, the region is expanded to 1 Mb upstream and downstream as this is a distance where most gene-enhancer complexes are included. (McLean et al., 2010) FUMA is a broader tool that uses GWAS summary statistics to find causal SNPs, functionally annotates them and prioritizes genes. It uses multiple sources to prioritize genes associated with causal SNPs, one of which is positional mapping. In this approach, a standard setting of 10 kb around each gene is applied. (Watanabe et al., 2017) Both tools allow the user to manually change these settings.

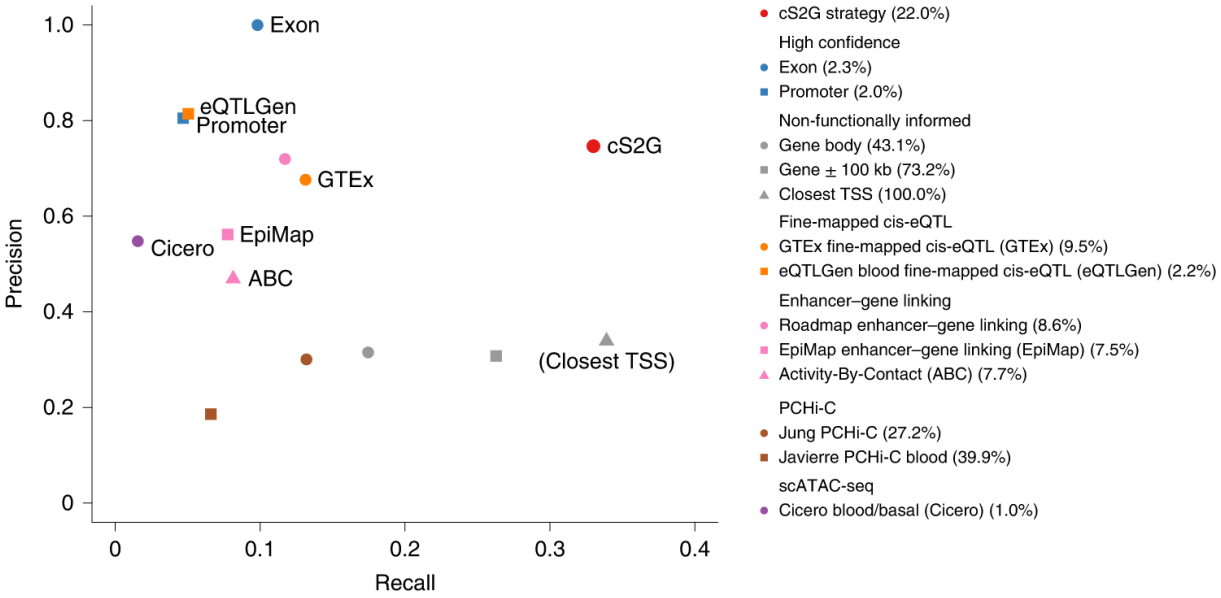


Figure 5: Precision-recall plot of different SNP-to-gene annotation strategies. *(Gazal et al., 2022) Grey dots represent distance-based strategies. Colored dots represent other strategies that use a functional foundation. The red dot represents the combination of strategies that have a functional foundation. Numbers mentioned for the different strategies in parentheses signify the percentages of SNPs that have minimum one gene linked to them.*

1.5.2 Gene expression-based methods

The main limitation of distance-based strategies where they do not consider any functional link between gene and locus missing distal gene-enhancer interactions, is followed up by gene expression-based strategies. These strategies use expression Quantitative Trait Loci (eQTLs) which are loci containing genetic variants like SNPs that are statistically associated to gene expression. (Umans et al., 2021) A difference is made between cis-eQTLs and trans-eQTLs. Cis-eQTLs are loci located close to the gene that is expressed. (approximately <1Mb). Trans-eQTLs, on the other hand, are further away from the expression signal (>5Mb). (Võsa et al., 2021) They have smaller effect sizes than cis-eQTLs which makes it more difficult to detect them in smaller sample sizes. These weaker effects make it less likely for trans-eQTLs to be compensated for its expression levels with post-transcriptional buffering. Also, strong cis-eQTL effects may have a big impact on an organism's fitness which increases the probability of getting removed by negative selection. Therefore, the collective impact of trans-eQTLs may be more significant for complex traits as it will not be compensated for or undergo negative selection. (Lee, 2022)

Investigating whether a GWAS signal and an eQTL signal share a causal variant, can result in a SNP being linked to a complex trait and to a significant expression level for a specific gene which could imply that the gene is linked to that trait. This phenomenon is called co-localization and forms a potential solution for linking genes to SNPs located in non-coding regions. (Zhu et al., 2016) The simplest way to find out whether GWAS loci and eQTLs are colocalized is doing a simple look-up and checking if they share the same variant on the same location. This method assumes that a single SNP affects both the phenotype and the gene expression level suggesting that this gene influences the phenotype. However, correlation does not imply causality. SNPs are often in LD with other SNPs which implies that the same SNP can be associated to a gene's expression and to a phenotype without the change in expression causing the phenotype. (Heinig, 2018)

A Bayesian statistical framework offers a powerful approach to assess whether a GWAS and an eQTL signal share a causal variant. This method is implemented in widely used tools such as COLOC and eCAVIAR. It combines prior beliefs with observed data—in this case, GWAS and eQTL summary statistics—to estimate posterior probabilities of shared causality. Each SNP is initially assigned an equal prior probability of being causal for either the phenotype or gene expression, and a lower prior probability of being causal for both, reflecting the assumption that shared causality is less likely. The framework assumes independence between the GWAS and eQTL signals, allowing the joint posterior probability of shared causality to be derived by multiplying the individual probabilities of each SNP being causal in the two traits. (Giambartolomei et al., 2014) Rather than evaluating SNPs in isolation, the Bayesian approach considers the entire pattern of association within a locus, accounting for LD and uncertainty in identifying the true causal variant. The framework can be further expanded because a locus can also contain multiple causal variants for the same phenotype or expression. This way the co-localization of a causal variant for the phenotype and a different causal variant for the expression level in the same locus can be detected reducing false negatives. (Hormozdiari et al., 2016) A key strength of this method is its probabilistic nature: it assesses the likelihood that each SNP is causal for one or both traits. Additionally, the use of priors enables the incorporation of biological knowledge or assumptions—such as the rarity of shared causality—by adjusting the weight given to different hypotheses accordingly. (Giambartolomei et al., 2014) In Figure 5, gene expression–based methods demonstrate high precision but low recall. This indicates that while the annotations identified through eQTLs are often accurate, these methods capture only a small fraction of all the true annotations that should be detected for the GWAS loci.

The usage of eQTLs has several important shortcomings that should be acknowledged. Firstly, eQTLs oversimplify the complexity of gene regulation. Gene expression is rarely regulated by a single eQTL. Instead, it results from a complex cascade of regulatory events occurring at multiple levels of gene expression. The gene products produced by the expression of one gene can, in turn, influence the expression of other genes, forming a hierarchical and interconnected network. As a result, a gene identified by an eQTL may be reacting to upstream regulatory effects rather than being the primary driver of the phenotype. Secondly, gene expression is context-dependent and can vary across different biological conditions. A particular genetic variant may affect expression levels only in a specific cell type or tissue. Consequently, if gene expression is measured in bulk tissue or in a biologically irrelevant cell type, key eQTLs may remain undetected. In addition to this spatial variability, temporal variability also plays a significant role. Gene expression can change across developmental stages or in response to environmental or physiological conditions. Some eQTL effects are therefore only detectable at specific time points or under certain circumstances. (Lee, 2022) The lack of data from diverse tissues or developmental stages further expands this problem, rendering many eQTL studies underpowered, particularly when key tissues or timepoints are not well-represented in the datasets. (Heinig, 2018) These layers of complexity pose significant challenges for the reliable identification and interpretation of eQTLs. Other mechanisms like epistasis which is the effect of one genetic variant being influenced by the presence of another variant and population stratification which shows that different populations have different effects, play a key role in the accuracy of eQTL mapping. (Gruber, 2007; Wei et al., 2014)

1.6 Gene Ontology Enrichment Analysis

Gene Ontology Enrichment Analysis (GOEA) is a method for functional annotation of a list of genes which are descriptions for cellular components, biological processes and molecular functions that define genes. The descriptions are called GO terms and are interconnected with each other in a hierarchical structure going from general to very specific concepts represented in a Directional Acyclic Graph (DAG). GO terms describe and thus can be associated to genes which are stored in GO annotations. GO itself is the framework that stores these descriptions and interconnects them through the annotations. GOEA tests which GO terms are overrepresented in an input set of genes. (Tomczak et al., 2018)

GOEA is widely used for interpreting the biological meaning of large-scale genetic data. The GO terms and annotations constantly grow as more data becomes available through experiments. Currently, GO consists of information from over 18 000 scientific papers with over 1 000 000 experimentally defined annotations. Annotations not originating from experimental data are called “electronic” annotations which are automatically generated GO annotations based on computational predictions. (Tomczak et al., 2018) Currently, over 2 400 000 annotations are electronic. (*Gene Ontology Resource*, n.d.) Despite the mix of experimental and electronic annotations, GO remains a widely used and trusted resource for functional interpretation, continually evolving as new data and improved methods become available.

1.7 Aim: novel gene annotation strategy

This thesis aims to overcome the limitations of distance-based methods that do not have a functional foundation and the limitations of gene expression-based methods which rely on eQTLs by developing a novel method for gene annotation that does have a functional foundation and does not need to rely on eQTLs. Expectation-Maximization for the Annotation of genes using GO enrichment (E-MAGO) initially annotates genes to GWAS loci based on proximity. By incorporating TADs, the pool of genes that could be annotated to a GWAS locus is limited to the TAD overlapping with the locus. This way the search space is limited to regions where genomic elements frequently work together. Then, using an Expectation-Maximization (EM) algorithm where the expectation step involves GO enrichment analysis to score genes and the maximization step selects the highest scoring genes, the genes annotated to the SNPs are iteratively updated until convergence.

The idea behind E-MAGO is to use a set of known genes to connect biological processes to them which are likely to be relevant to the phenotype. Having this biological context, the search space can be more focused on the relevant genes for these processes and prioritize genes according to their functional relevance.

In a next step, E-MAGO will be compared to the distance-based method, GREAT using examples of well-studied complex traits and diseases. First, results will be evaluated qualitatively by looking into curated databases which provide genes for diseases and traits with high confidence. Subsequently, the precision of gene annotation will be benchmarked against eQTL-GWAS co-localization results, which will serve as the ground truth due to their established accuracy.

2 Materials and Methods

2.1 Theory

2.1.1 The Expectation-Maximization algorithm

The EM algorithm is an iterative method that searches for the maximum likelihood estimate of parameters given some observed data. The EM algorithm has the advantage of being able to estimate parameters even when some data is missing or inaccessible, treating this unavailable information as hidden variables. (Do & Batzoglou, 2008) To get an intuition on this algorithm a toy example is given. Imagine having two mixtures with blue and green marbles from which 2 marbles are drawn. It is unknown from which mixture each marble is drawn. To find out the distribution of marbles in each mixture, the EM algorithm will compute the probability that each marble came from each mixture based on an initial guess of distributions in the mixture. Then, using these probabilities the current distributions are updated to maximize the likelihood of observing the drawn marbles in the distributions. These steps are executed iteratively until the distributions do not change anymore.

The EM algorithm is divided into two steps: the expectation and the maximization step. The algorithm starts with an initial guess of the parameters Θ that it tries to estimate. The observed data can be denoted as $Y = (y_{1,}\ y_2, \dots, y_i)$ with i the number of observations and the hidden data as $X = (x_1, x_2, \dots, x_j,)$ with j the length of the hidden data. (Moon, 1996) Now, the idea is to find the maximum log likelihood estimate of Θ given the observed data as illustrated in Equation 1. (Dierckx & Moreau, 2024)

$$\Theta^{ML} = argmax_{\Theta} \ln P(Y|\Theta) \quad (1)$$

The log-likelihood is computed instead of the likelihood as values can get small and highly skewed. This way values can be more numerically stable. Also, transforming helps with handling non-linearity, making it easier to do optimization. (Benoit, 2011) Calculating the maximum likelihood means knowing the likelihood, but the likelihood is not only based on Y, but also on X. Since X is unobserved, the log-likelihood cannot be calculated. That is where the expectation step of the EM algorithm comes in play. This step computes the expectation of the log-likelihood with X and Y over the distribution of X and given Y and the current parameters Θ. (Dierckx & Moreau, 2024) The expectation step is presented in Equation 2.

$$\ln P(Y|\Theta) = E_{X|Y,\Theta}(\ln P(X, Y|\Theta)) \quad (2)$$

This expectation is often written as the auxiliary function $Q(\Theta|\Theta_k)$ where k indicates the current iteration. The subsequent maximization step then maximizes this expected log-likelihood to compute the parameter estimates for the next iteration $k + 1$ illustrated in Equation 3. (Dierckx & Moreau, 2024)

$$\Theta_{k+1} = argmax_{\Theta}\ Q(\Theta|\Theta_k) \quad (3)$$

Finally, convergence is reached when the parameter estimates at iteration k and k+1 are equal or under a specific threshold implying that that estimates do not change anymore. The result is the maximum likelihood estimate of Θ given the observed and hidden data. A visual representation of this can be found in Figure 6. In this figure the original likelihood $P(Y|\Theta)$ is shown. As the iterations increase the auxiliary function moves closer to the optimum. Equation 4 illustrates the interplay between the expectation and maximization steps with Θ_k the resulting optimal parameter estimates where convergence was reached after k iterations. (Moon, 1996) The tangent becomes horizontal at the optimum which is also an indicator of convergence.

$$\Theta_0 \xrightarrow{E} \ln P(Y|\Theta_0) \xrightarrow{M} \Theta_1 \xrightarrow{E} \ln P(Y|\Theta_1) \xrightarrow{M} \Theta_2 \xrightarrow{E} \ln P(Y|\Theta_2) \xrightarrow{M} \dots \xrightarrow{M} \Theta_k \quad (4)$$

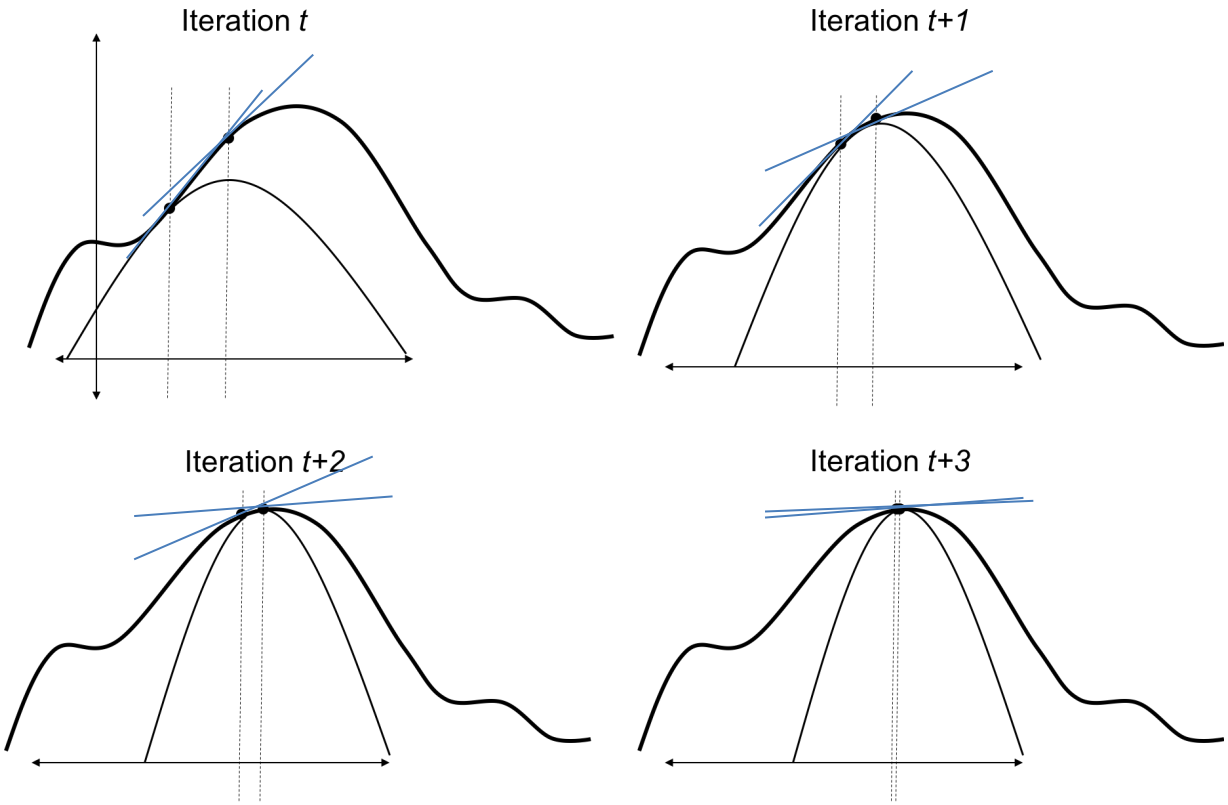


Figure 6: Graphical representation of the likelihood function. *(Bernstein, 2020) The X-axis represents Θ and the Y-axis represents the likelihood of the estimates at iteration t, the thick curve represents the likelihood for different Θ's and the thin curve represents the auxiliary function which changes at every iteration as the estimates change, the blue lines represent the tangent of the auxiliary function.*

Note that if the first guesses for the parameter estimates were different, the auxiliary function could have started at another location on the likelihood curve where it could have found other optima or get stuck in saddle points. This implies that it is not always guaranteed to find the global maximum. A solution to find global convergence would be to do random restarts with different random first guesses of the parameters. Out of the maximum likelihood results, the one with the highest likelihood has the highest chance of being the global maximum. Instead of random guesses, prior knowledge could help

to state prior beliefs about the initial parameter estimates and consequently finding the global optimum. (Ng, 2013)

Nevertheless, the EM algorithm is widely used and conceptually straightforward. It is an efficient method capable of handling analytically complex problems where directly computing the optimum is not feasible. With appropriate data or initialization strategies, EM can also be robust in converging to a global optimum.

2.2 Implementation

E-MAGO, a new gene annotation strategy based on the EM algorithm was developed in Python 3.11. (Salgado, 2022) The observations are the lead SNPs from a GWAS study of choice, the hidden variables are the true target genes of the SNPs, and the unknown parameters are the scores given to the genes that rate their likelihood of being the target of any SNPs. These scores are updated in every iteration.

2.2.1 Data organization

All the data used in E-MAGO whether it is imported or generated, is linked to each other. To have a clear overview and to keep the data organized a database was created in which the data is stored in different tables that are interconnected. The database was created using MySQL which is a widely used open-source SQL database management system. (Letkowski, 2015)

In total, six tables were created. In Figure 7 a schematic overview of the different tables and their columns is presented. The primary key of a table is represented by underlining the columns participating in the primary key. The primary key represents a column or a combination of columns that is unique for every row in the table. Black tables are the tables that are predefined and stay the same across different runs of the method. Blue tables are the tables that start empty at the beginning of the execution of the method and that are filled in according to the input of the user.

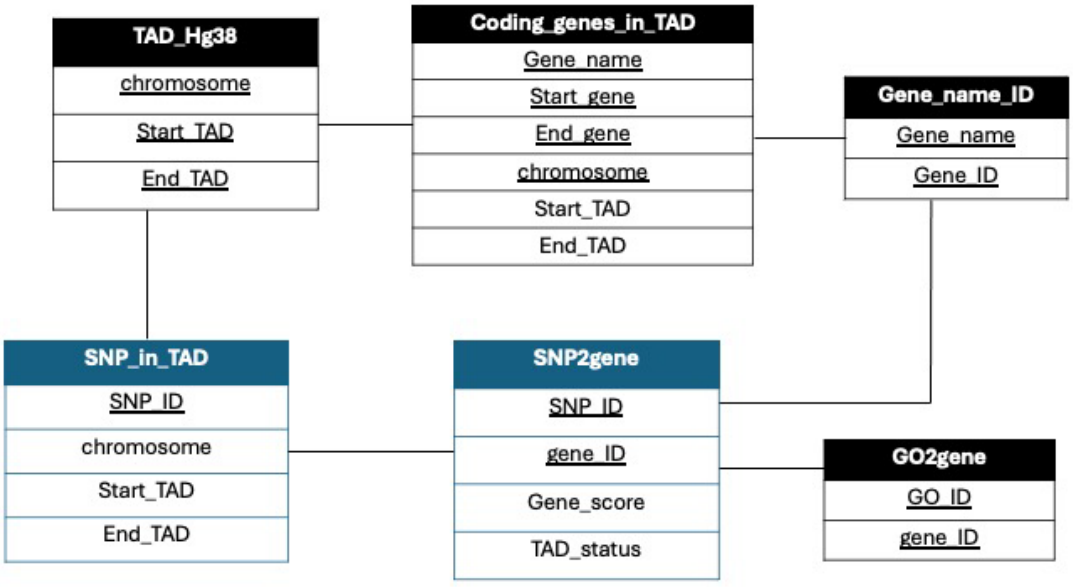


Figure 7: Schematic overview of E-MAGO's database design.

The “TAD_Hg38” table consists of the TAD data which is defined as the different TAD boundaries which represent the different TADs on the chromosomes. The TAD boundaries were obtained from the Knowledge Base of Topologically Associating Domains (TADKB). (T. Liu et al., 2019) In TADKB, the TAD boundaries are found by analyzing Hi-C data and taking the overlapping TADs found in 2 clustering methods. The first method is structural clustering which is based on 3D genome clusters. The second one is chromatin-state clustering where the similarity in functional landscapes is compared. Manual validation was also done by comparing TADKB’s boundaries with Hi-C maps from the UCSC genome browser. (Karolchik et al., 2003) The overlap between the boundaries and high contact frequency regions were checked. (Supplementary Data 6.1)

The ”Coding_genes_in_TAD” table stores all the coding genes and assigns them to the TAD they are located in. Coding genes were fetched from the UCSC genome browser through their open-source relational database. Coding genes outside the TADs are not included in this table.

The “Gene_name_ID” table is used to convert from gene name to its gene ID. This is necessary to keep consistency and connectivity as for example the background genes from the Gene Ontology Consortium (GOC) come in a list with gene IDs, but fetching protein coding genes from the UCSC genome browser is done with gene names.

The “GO2gene” table represents the connections between GO IDs and gene IDs where genes participate in the GO term of the given GO ID. This table helps with fast look-up to connect gene IDs with GO IDs without having to load in the full GO DAG every iteration.

2.2.2 User Input

The input provided by the user is the lead SNPs from a GWAS study usually in a bed format which is frequently used to represent genomic features per line. (Quinlan & Hall, 2010) The file has 4 tab-separated columns. The first column consists of the chromosome number, the second column is the position of the SNP minus one, the third column is the position of the SNP, and the fourth column contains the SNP identifier. Any type of SNP identifier is allowed. The most used one is the reference SNP cluster ID (rsID) which originates from dbSNP. (Sherry et al., 2001)

Important to note is that all the data in the method use the GRCh38 reference genome as this is the most recent and most accurate human reference genome. (Guo et al., 2017) Most studies use this reference genome, but when the input of the user is based on another reference like GRCh37 the LiftOver tool from the UCSC genome browser can be used to convert coordinates to align with GRCh38.

2.2.3 Initialization

The method starts with linking every SNP to a TAD in which it resides. This way the search space is limited and by using a TAD it is ensured that all genomic elements are functionally linked. When a SNP resides between two TADs, its search space becomes the inter-TAD region and both TADs. If a SNP is located just upstream of the first TAD or just downstream of the last TAD on the chromosome, it is assigned to that first or last TAD, respectively. This information is stored in the “SNP_in_TAD” table.

Having the “Coding_genes_in_TAD” and the “SNP_in_TAD” table every SNP can now be linked to the protein-coding genes situated inside the same TAD which is stored inside the “SNP2gene” table. Some TADs happen to be empty which results in SNPs that do reside in that TAD to have no genes linked. In those edge cases the SNP will be linked to any genes 1000 bp up- or downstream. The “SNP2gene”

table has a column “TAD_status” which will be equal to one if the SNP is linked to genes it found in its TAD, otherwise it will be equal to 0.

All SNPs are now linked to protein-coding genes and therefore every SNP has a set of candidate causal genes. This whole procedure is visualized in a flowchart in Figure 8. The first step in the EM algorithm is to initialize the parameter estimates. In E-MAGO, this involves assigning the gene closest to each SNP as the initial estimate of the likely target gene.

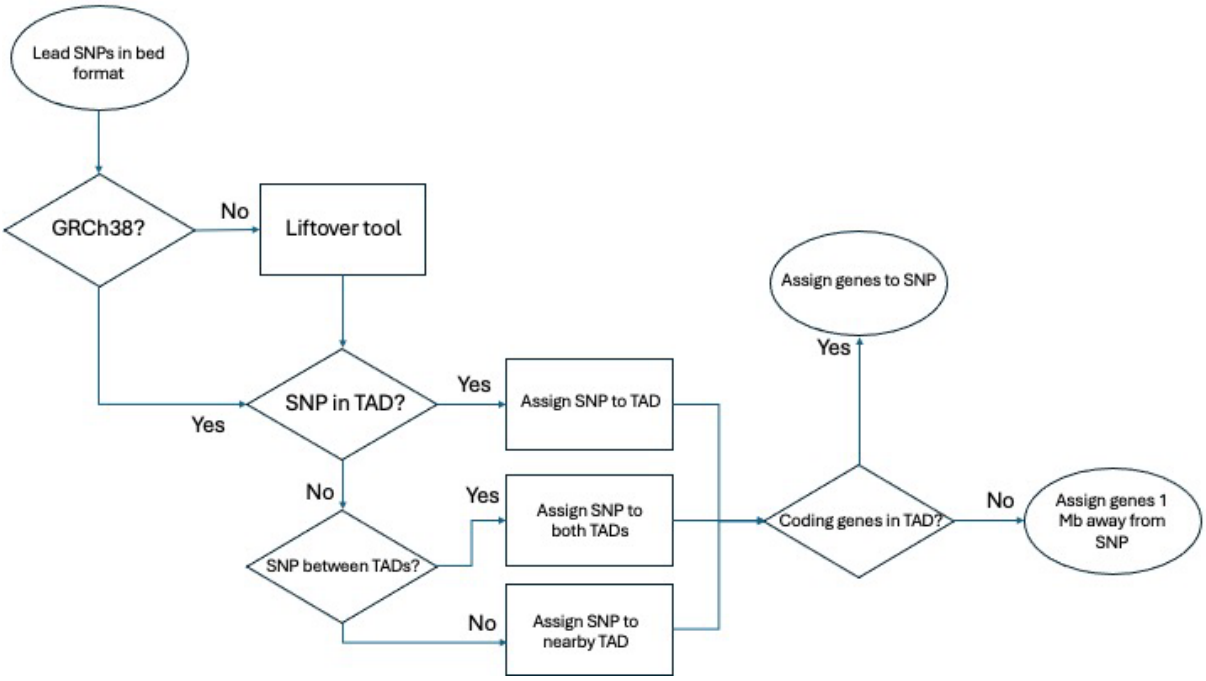


Figure 8: Flowchart of the data preparation in E-MAGO.

2.2.4 Expectation

The EM algorithm defines the expectation step as finding the expected log likelihood of the data including the hidden variables. In other words, calculating the likelihood of the complete data given the current parameter estimates. In the novel method this translates as calculating the likelihood that each gene is the correct causal gene for a given SNP given the gene scores calculated using GOEA.

$$\ln P(input\ SNPs|gene\ scores\ per\ SNP) = E_{causal\ genes|gene\ scores,input\ SNPs}(\ln P(input\ SNPs, causal\ genes|gene\ scores)) \quad (5)$$

Concretely, every iteration GOEA is executed on the set of candidate genes that were chosen at the maximization step or if it is the first iteration, on the genes chosen at initialization.

GO enrichment is done using the Python-based library, GOATOOLS (Gene Ontology Analyses tools). (Klopfenstein et al., 2018) This tool does GOEA using the Fisher’s exact test, which is a statistical association test for two binary, categorical variables in a contingency table. The test evaluates the different combinations of values of the variables and computes whether the variables are associated or not. (Preacher & Briggs, 2001) For GOEA this means testing if genes in the input set are more likely to have the GO term annotated versus genes not in the input set. The probability that is calculated is

based on a hypergeometric distribution which gives the probability distribution of having a or more overlapping genes assuming random sampling and without replacement. (Wu, 1993)

Table 1: Contingency table for GOEA using Fisher's exact test.

	Genes annotated with GO term	**Genes not annotated with GO term**
Genes in input set	a	b
Genes not in input set	c	d

In GOEA only overrepresentation is of interest which means only a right-sided test is done and only the right tail of the hypergeometric distribution is investigated. For each GO term separately, the test calculates the probability of observing the values in Table 1 by chance or when there is no overrepresentation of the GO term in the input set.

$$P = \sum_{i=a}^{min\,(a+c,a+b)} \frac{\binom{a+b}{i}\binom{c+d}{a+c-i}}{\binom{a+b+c+d}{a+c}} \tag{6}$$

$$H_0 = no\ overrepresentation \tag{7}$$

$$H_A = overrepresentation \tag{8}$$

Equation 6 represents the formula for calculating this probability. (Preacher & Briggs, 2001) The sum adds up all probabilities for i or more overlapping genes which results in the probability of getting i or more overlapping genes by chance. It goes from a which are the genes in the input set and annotated with the GO term until the minimum of $a + c$ which are all genes annotated with the GO term and $a + b$ which are all genes in the input set. This minimum represents the maximum possible overlap as there cannot be more genes than the ones in the input set or annotated with the GO term that overlap. The denominator counts all the ways to arrange the genes while keeping the total count. The first term in the numerator represents all the ways to choose i from the input set and the second term represents all the ways to choose $a + c - i$ from the genes not in the input set.

The GO has over 45 000 terms and 134 000 connections between these terms. This implies that GOEA must do an enormous number of hypothesis tests which will cause a lot of false positive and false negative results. (Carbon et al., 2019) In E-MAGO, GOATOOLS uses the Benjamini-Hochberg False discovery Rate (FDR). The FDR represents the proportion of false discoveries. For example, an FDR of 5% means that 5% of the statistically significant results will be false. Some false positives are allowed, but it controls that this stays around 5%. In GOEA, this is accepted, because it is more about finding what the biological context is for a group of genes and not about every single result being correct. That

way FDR can capture more true positive results than other stricter correction methods. (Goeman & Solari, 2014)

To execute GOEA in GOATOOLS a GO DAG and GO annotations are needed. The GO DAG file, *go-basic.obo* is downloaded from the GO website (https://geneontology.org/docs/download-ontology/). The GO annotations file called *gene2go* is downloaded from the GOC (https://ftp.ncbi.nlm.nih.gov/gene/DATA/). Further, a set of background genes is needed for GOEA to compare the input genes with. A list of genes captured by the GOC is used which were filtered with amiGO2 for "homo sapiens" and "protein".

From the GOEA output, GO terms with a corrected p-value $\leq$0.05 were retained. In case less than 10 GO terms are significant which is more likely for gene sets that are small, then the top 10 GO terms are used, because that way enough biologically relevant information can be used for scoring genes. All genes in the search space annotated to a significant GO term get a score based on the significance level of the GO term. Every time a gene is linked to a significant GO term the gene score increases based on its significance level.

$$gene\ score = \sum_{i} \log_{10} corrected\ pvalue_i$$

(9)

With i equal to the significant GO terms linked to that gene. Now, every input SNP that has one of the genes that has been scored, will now update the gene score for that gene inside their search space by adding up the gene score found in the current iteration. The gene scores for the genes linked to the SNPs represent the likelihood that each gene is the target gene for that SNP. Naturally, a higher score means a higher likelihood.

2.2.5 Maximization

The EM algorithm defines the maximization step as maximizing the expected log likelihood. This implies taking the gene with the highest score and thus the best parameter estimate for each SNP and using that gene as input for the next iteration.

$$gene\ scores_{k+1} = argmax_{\text{gene scores}}\ Q(gene\ scores|gene\ scores_k)$$

(10)

Equation 10 shows the updated gene scores at iteration $k + 1$. The genes with these maximum gene scores for each SNP are then used in the subsequent expectation step.

2.2.6 Convergence

The expectation and maximization steps are repeatedly executed. Iterations stop at convergence which is when the genes assigned to the SNPs in previous iteration are exactly the same genes assigned to the SNPs in the current iteration. This implies that after updating gene scores the same genes still have the best gene scores.

$$P(target\ genes|gene\ scores_{k+1}) = P(target\ genes|gene\ scores_k)$$

(11)

2.3 Disease-linked genes

GeneCards is a database for human genes. It is a part of the GeneCards Suite which is a large knowledgebase containing different human databases including MalaCards which is a human disease database and PathCards which is a database for human biological pathways. All these databases are presented as a web-based platforms making it user-friendly. By inputting the disease or trait of interest, GeneCards will search through over 150 sources to find which genes are associated to the input disease or trait and order them based on a relevance score. The scoring is done using ElasticSearch which is a search engine that collects, processes and analyzes enormous amounts of data. It can search through any document for a term and assign it a relevance score based on how well it matches. The matching process calculates the term frequency in a document while giving more weights to terms that are rare over all documents and less weight to terms that appear in longer documents. In GeneCards the terms are diseases. The documents are resources with gene entries that go from genomics found through HGNC, Ensembl, NCBI and UCSC Genome Browser, expression data found through GTEx, proteomics from UniprotKB and InterPro, scientific publications from PubMed and various other sources. Genecards adds extra weights to different sections that are more relevant. For example, the section “disorders” would be of higher importance than “functions”. (Stelzer et al., 2016)

2.4 GO enrichment with g:Profiler

The biological relevance of the resulting gene sets was evaluated using GOEA via g:GOSt, a feature of the g:Profiler web-based platform. (Raudvere et al., 2019) This tool performs functional enrichment using Fisher’s one-sided exact test to identify GO terms that are statistically overrepresented in the input gene list.

g:Profiler is user-friendly and requires only a gene list to perform enrichment analysis. Its underlying background database is regularly updated to reflect current biological knowledge. A significance threshold of $p < 0.05$ is applied, along with g:SCS, g:Profiler’s custom method for multiple testing correction. The g:SCS algorithm sets a global p-value threshold based on the size of the gene set and the degree of overlap among GO terms, recognizing that many GO categories share common genes.

2.5 GWAS data

Experiments were conducted on different GWASs found through the GWAS catalog namely on Breast Cancer (BC), Inflammatory Bowel Disease (IBD), Type 2 Diabetes (T2D), Schizophrenia and Alzheimer’s Disease. For the BC study a large-scale meta-analysis was performed combining GWAS data from over 300 000 individuals of Asian and European ancestry. Comprehensive Quality Control (QC) steps were applied to ensure high data accuracy, including filtering based on imputation quality, MAF and population stratification. Additionally, replication of findings in an independent cohort strengthened the validity of the discovered loci. (Shu et al., 2020) The IBD study analyzed data from 25 305 individuals and conducted a meta-analysis with published summary statistics, resulting in a total sample size of 59 957 subjects. The study's large cohort enhances its statistical power to detect genetic associations. (De Lange et al., 2017) The T2D study analyzes genetic data from 421 743 individuals with 51 256 T2D cases and 370,487 controls coming from six ancestry groups across five cohorts, including large biobanks like UK Biobank, BioMe, and the Mass General Brigham Biobank. The study applied various QC steps including filtering based on imputation quality, MAF, and population stratification. (Huerta-Chagoya et

al., 2024) These three diseases are extensively studied and supported by well-powered GWAS datasets, making them strong candidates for evaluating and comparing the performance of different annotation methods. Additionally, Schizophrenia and Alzheimer's disease GWASs were used to add extra power to one of the experiments. For Schizophrenia, the study involved a sample size of 108 341 individuals, comprising 43 175 cases and 65 166 controls. QC procedures consisted of standard filtering steps based on imputation quality, MAF and population stratification. (Z. Li et al., 2017) Regarding Alzheimer's Disease, the study used 71 880 cases and 644 457 controls. Standard filtering steps were conducted and an extra QC step was conducted to account for inflation due to polygenicity and confounding. (Adewuyi et al., 2022) Incorporating these studies into the experiment provided additional power to evaluate and compare the performance of different annotation methods across a diverse set of diseases, thus strengthening the overall findings.

3 Results

3.1 Annotating genes to Breast Cancer GWAS loci, an example

The first step of E-MAGO involves assigning each breast cancer (BC) lead SNP to a TAD. Lead SNPs located within or on the boundaries of a TAD are assigned to that TAD. If a lead SNP lies between two TADs, it is linked to both. Similarly, SNPs located just before the first TAD or just after the last TAD are assigned to the nearest TAD. In total, the BC GWAS study identified 148 lead SNPs. Of these, 140 were located within a TAD, 8 were situated between two TADs, and none were found outside the TAD boundaries entirely. All genes located within the same TAD as a given SNP are considered candidate genes for that SNP. The number of genes per TAD can vary widely — some TADs are gene-rich, while others may contain no genes at all. When a SNP falls within a gene-empty TAD, genes located within 1 Mb upstream or downstream of the SNP are used instead. Figure 9 illustrates the distribution of candidate gene counts per SNP. On the left, SNPs within TADs are shown; on the right, SNPs without associated TAD genes are shown. Most TAD-linked SNPs are associated with 5 to 10 genes, although some SNPs, located in gene-dense regions, are linked to over 50 genes. Only a few SNPs had no genes within their TADs. For these, applying the ±1 Mb window resulted in 4 to 13 candidate genes.

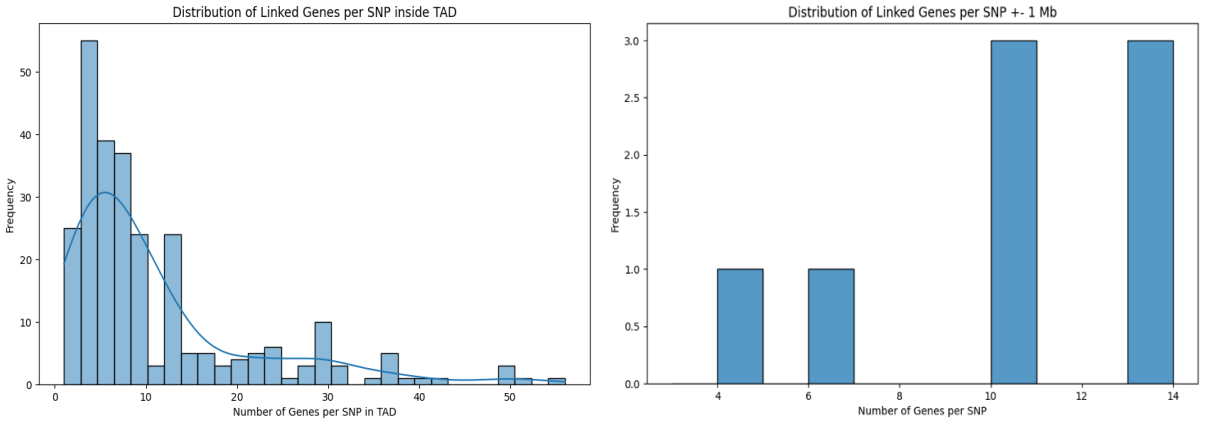


Figure 9: The distribution of the genes. *The left barplot shows the distribution of the number of genes that are linked per SNP with on the X-axis the number of genes that are in the TAD of the SNP and on the Y-axis the number of SNPs that have that number of genes in their TAD. The right barplot shows the number of genes assigned to a SNP when the SNP does not contain any genes in its TAD with on the X-axis the number of genes per SNP and on the Y-axis the number of SNPs that have that number of genes.*

The initial genes used for GO enrichment are those most closely linked to each SNP. GO enrichment analysis identifies significantly overrepresented biological processes among these genes. Each time a gene is associated with an enriched GO term, its gene score is updated to reflect its potential functional relevance. Let's take the lead SNP rs71338792, which is linked to 30 genes. In Figure 10, the evolution of gene scores across iterations is shown. As the iterations progress, genes that are more frequently associated with enriched GO terms accumulate higher scores. This highlights their potential biological relevance, shifting the focus from merely proximal genes to those more functionally involved in the disease process.

Notably, FOSB consistently emerges as the gene most strongly associated with enriched terms, maintaining a high score throughout the iterations. In contrast, although SIX5 initially ranks highly as a nearby gene, it is overtaken by ERCC2 after two iterations. This shift reflects the functional importance of ERCC2—a gene involved in DNA repair—compared to SIX5, which is a transcription factor with less direct evidence of involvement in breast cancer pathways. The dynamic changes in ranking emphasize how this iterative method prioritizes genes with stronger biological relevance to the disease.

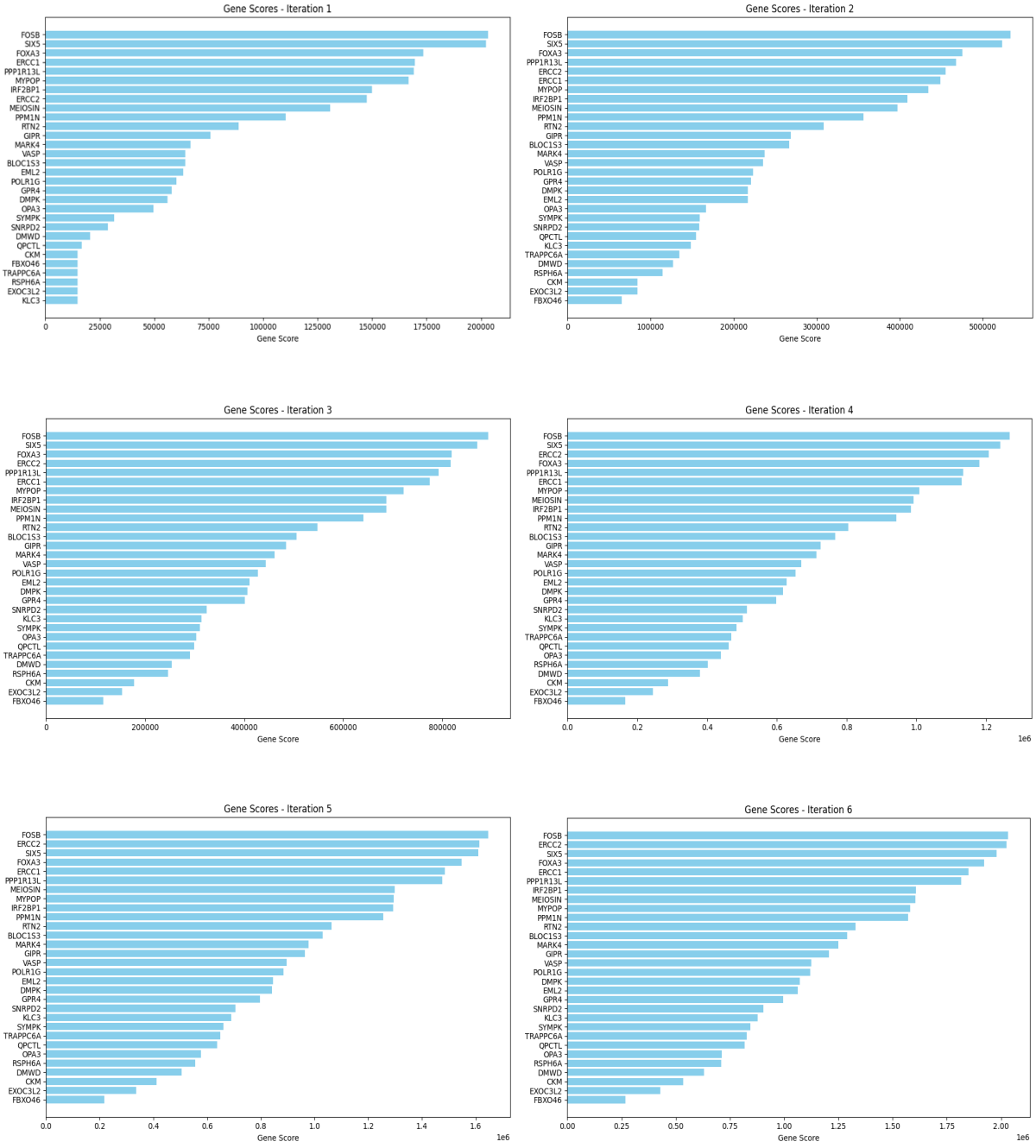


Figure 10: The evolution of gene scores through the iterations. *Illustrated through barplots with on the X-axis the gene score and on the Y-axis the genes ranked in decreasing order from top to bottom.*

In each iteration, the top-ranked gene for every SNP is included in the subsequent GO enrichment analysis. Genes that are consistently associated with significant GO terms accumulate higher scores, increasing their likelihood of being selected again in the next round. This creates a feedback loop in which biologically relevant genes become more prominent over time, and the GO terms they are linked to are more likely to reappear as enriched.

To visualize these dynamics, Figure 11 shows how the top enriched GO terms evolve across iterations. To enhance interpretability and reduce redundancy among terms, GO Slim mapping was applied. This groups detailed GO terms into broader, more general categories known as GO Slim terms, providing a clearer overview of the underlying biological processes.

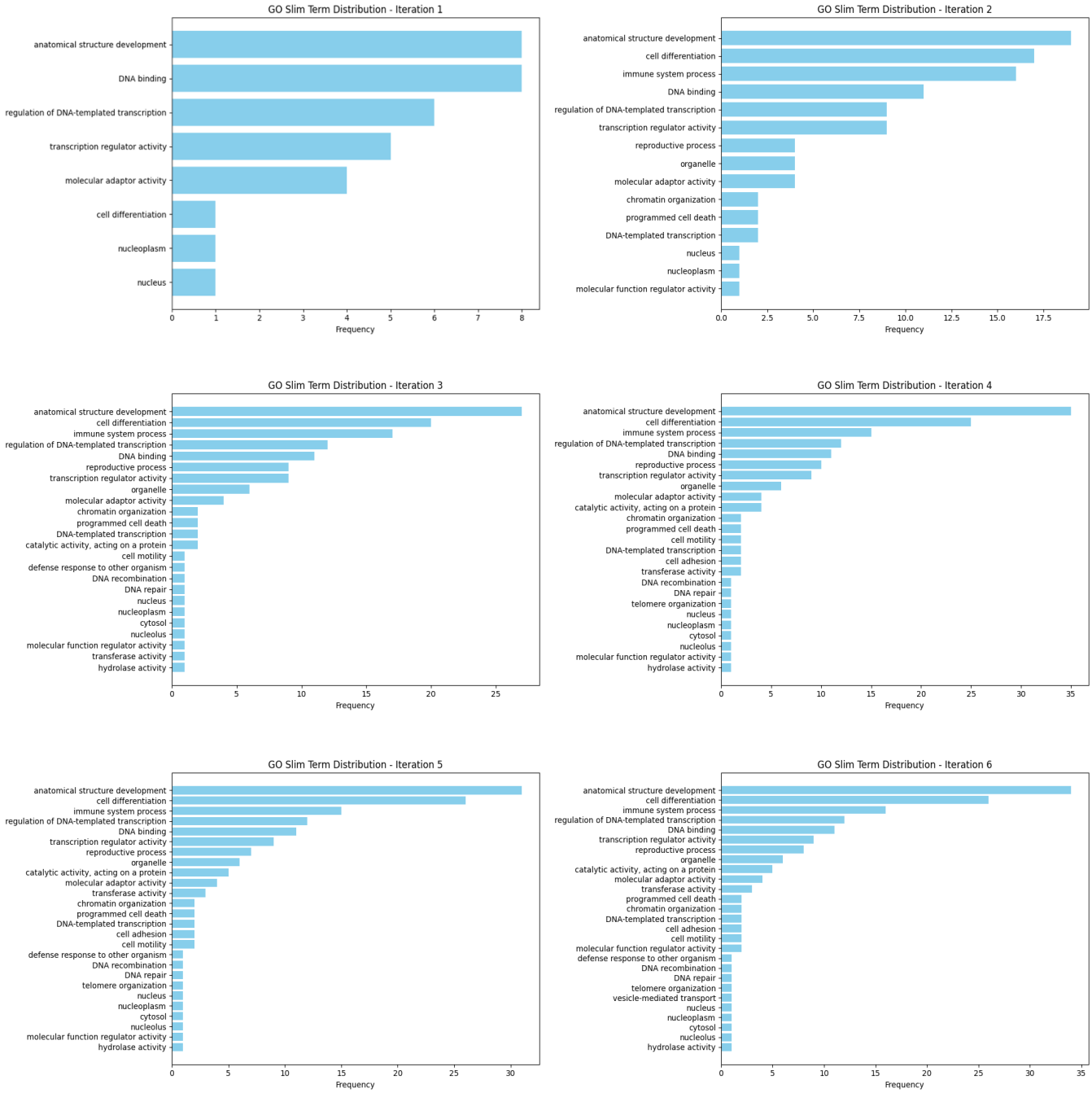


Figure 11: The evolution of GO slim terms through the iterations. *Illustrated with horizontal barplots with on the X-axis the number of times GO terms with categorized into a certain GO slim term and on the Y-axis the GO slim terms ranked in descending order.*

the term “anatomical structure development” is the most prominent term through all the iterations. It includes terms that are biological processes that shape any part of the body from cell development to organ or skeletal development. In the context of BC this includes angiogenesis, cell proliferation and breast-tissue development. The terms “cell differentiation” and “immune system process” become more prominent during the iterations which makes sense because cancer disrupts normal cell growth and immune responses. Also, other important BC terms are picked up like programmed cell death, DNA repair and telomere organization.

3.2 Validation of E-MAGO

The goal of these experiments is to validate the effectiveness of E-MAGO in identifying the target genes of lead SNPs. The performance of E-MAGO was compared against a commonly used distance-based annotation method, GREAT, to determine whether E-MAGO provides improved results. Specifically, candidate gene lists generated by E-MAGO for GWAS loci associated with BC, IBD, and T2D were compared to those produced by GREAT.

In the first experiment, both methods were evaluated against disease-associated gene sets retrieved from GeneCards, which served as the ground truth. The degree of overlap with GeneCards was assessed, and genes uniquely shared between E-MAGO and GeneCards—but missed by GREAT—were examined using the UCSC Genome Browser. This analysis aimed to determine whether GREAT’s reliance on nearest-gene mapping fails to capture functionally relevant genes in regions containing multiple candidate genes. GO enrichment analysis was then performed on the resulting gene sets to evaluate their biological relevance to the respective diseases. Lastly, precision and recall for E-MAGO and GREAT against GeneCards per disease were calculated. Precision signifies the fraction of predicted genes by E-MAGO or GREAT that were also captured by GeneCards. Recall is the fraction of all disease-linked genes found with GeneCards that were also found by E-MAGO or GREAT. These values show how well the different methods can capture relevant results even if the gene sets only capture very few positive results.

In a second experiment, eQTL-based data was used as ground truth to assess whether E-MAGO or GREAT more accurately identifies genes whose expression is affected by disease-associated variants. Similarly to the first experiment, the overlap between predicted genes and eQTL-supported genes was used to benchmark each method's performance. Precision and recall were calculated to evaluate the performance of the methods in a quantitative and uniform way.

3.2.1 Validation with GeneCards genes

3.2.1.1 Gene set evolution through iterations

GREAT annotates the gene that is in closest proximity to each SNP. Similarly, E-MAGO begins by assigning the closest gene as an initial guess. The overlap between the gene sets annotated by both methods was examined across multiple iterations. This allows for the observation of how the gene sets diverge as the novel method progressively scores genes based on GO enrichment. Figure 12 illustrates these changes over time.

For all three diseases, the similarity to the gene set identified by GREAT is highest at the initial step and decreases as the algorithm progresses. In the case of BC and T2D, the similarity levels off after two iterations, whereas for IBD, the overlap keeps decreasing slowly.

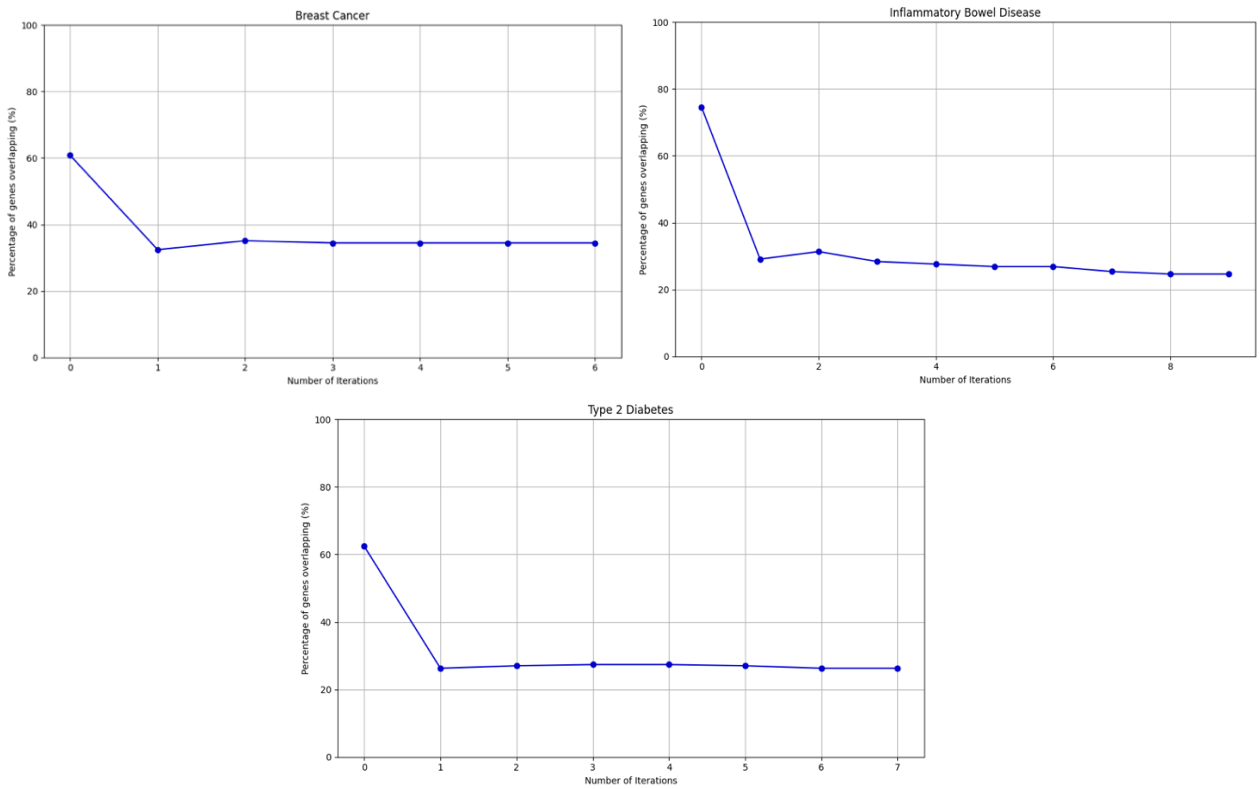


Figure 12: The change in similarity of the gene sets between GREAT and E-MAGO*. The x-axis represents the number of iterations E-MAGO has executed, and the Y-axis is the percentage overlap in gene sets between GREAT and E-MAGO.*

3.2.1.2 Overlap of genes

Using both E-MAGO and GREAT, candidate genes were annotated to 148, 134, and 274 lead SNPs from GWAS on BC, IBD, and T2D, respectively. For BC, E-MAGO and GREAT successfully annotated 147 and 148 SNPs, respectively, while GeneCards used its top 150 genes. For IBD, E-MAGO and GREAT both successfully annotated 134 genes and GeneCards used its top 150 genes. For T2D, E-MAGO and GREAT both successfully annotated 274 genes and GeneCards used its top 275 genes. Across all three GWAS datasets, E-MAGO was able to identify the same number or more disease-linked genes compared to GREAT (Figure 13). Distinct disease-linked genes were identified by E-MAGO and GREAT (Figure 14).

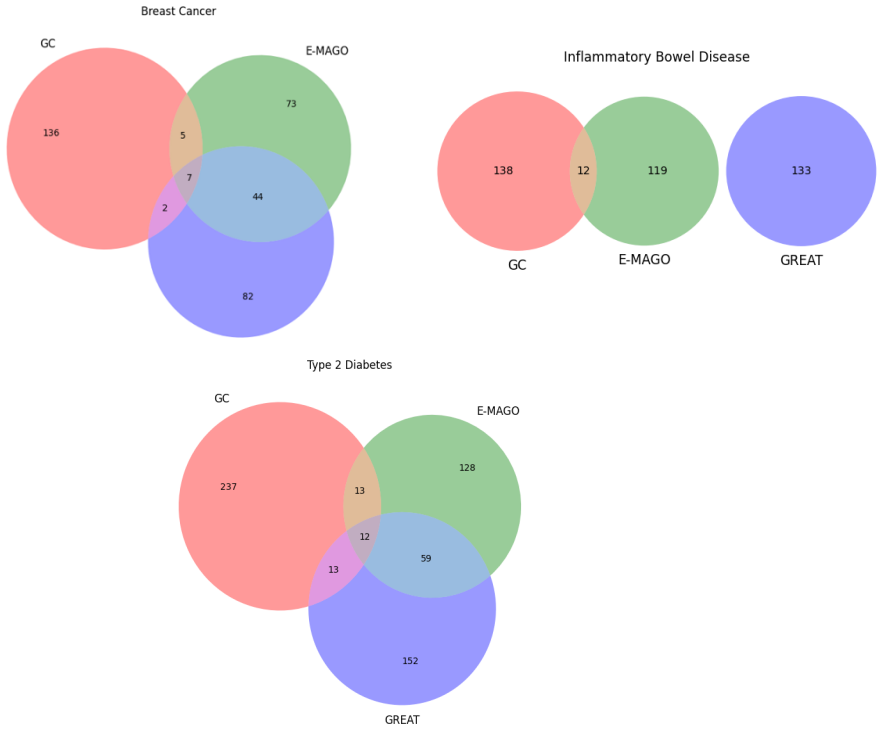


Figure 13: Venn diagrams illustrating the overlapping genes of GeneCards, E-MAGO and GREAT.

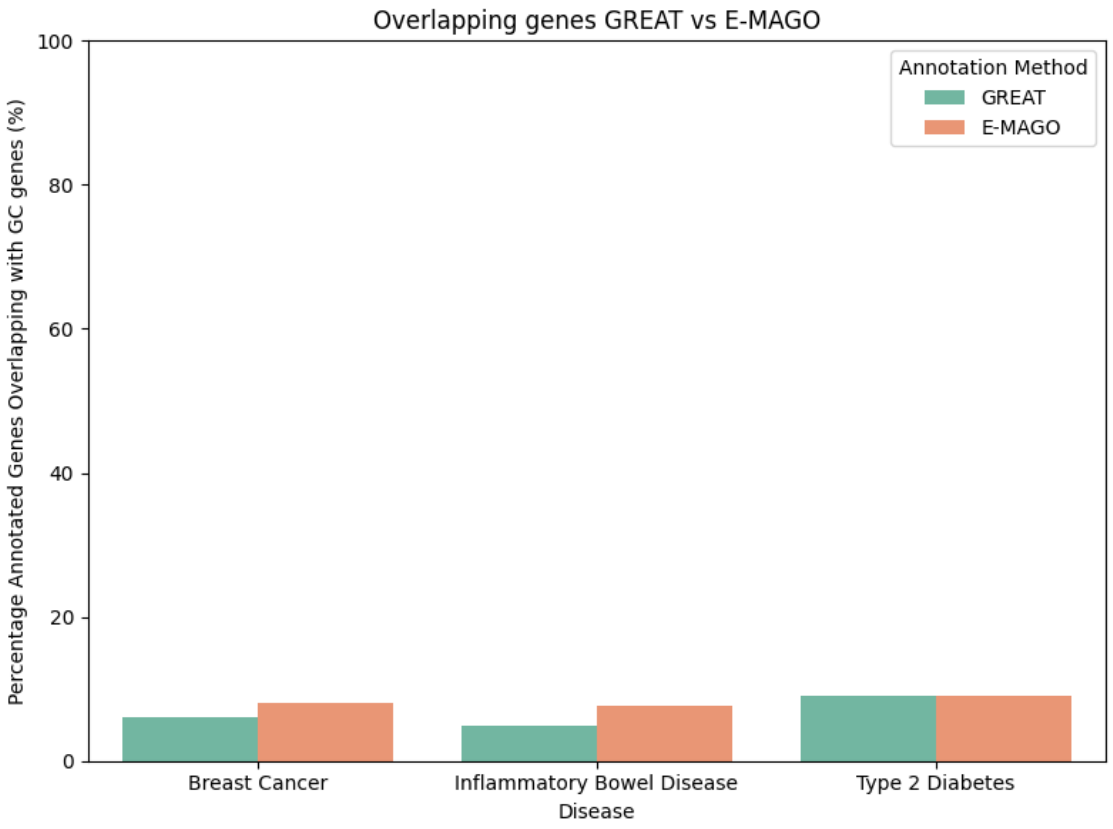


Figure 14: The comparison of gene overlap of E-MAGO and GREAT against GeneCards. *Illustrated with a paired barplot with on the X-axis the diseases and on the Y-axis the percentage overlap of genes between E-MAGO or GREAT and GeneCards.*

In TADs with many genes, the lead SNP is usually in close proximity to many potential target genes. Distance-based methods which only annotate the closest gene, are less likely to annotate the true target gene. Genes identified exclusively by E-MAGO and GeneCards were further examined to determine whether the associated SNPs are located in gene-dense regions. For every disease, one gene that was uniquely found by GeneCards and E-MAGO is illustrated inside its TAD (Figure 15). In Supplementary data 6.2, some extra examples are given. NF1, ITGB2 and GCK are examples of such genes for BC, IBD and T2D, respectively. Inside the TADs of NF1, ITGB2 and GCK are 7, 32 and 15 genes, respectively. The lead SNP of NF1 is located 190 kb upstream from NF1 while 5 other genes are located in a range of 70 kb from the lead SNP. More specifically, TEFM is the gene 3 kb from the SNP and annotated based on distance by GREAT. TEFM is a transcription elongation factor that regulates mitochondrial RNA polymerase activity which is not related to BC while NF1 negatively regulates the RAS signaling pathway which is important for cell proliferation and thus related to BC. The lead SNP of ITGB2 is located 690 kb upstream from ITGB2 while 7 genes are located in a range of 100 kb from the lead SNP. GREAT annotated ICOSLG which is 27 kb away and functions as a T-cell costimulator for T-cell proliferation. Immune processes are involved in IBD, but this gene does not have any evidence that it is involved in IBD. ITGB2 encodes the integrin beta chain which is part of the integrin that is important for cell adhesion meaning that it is involved in IBD. The lead SNP of GCK is located 47 kb away from GCK. There are 6 genes in a range of 80 kb from the SNP including GCK, but YKT6 is annotated to GREAT as it is 6 kb from the SNP. YKT6 involved in intracellular vesicle transport which is not directly linked to T2D. GCK is glucokinase which is crucial in glycolysis for triggering insulin secretion. Therefore, GCK is essential in T2D.

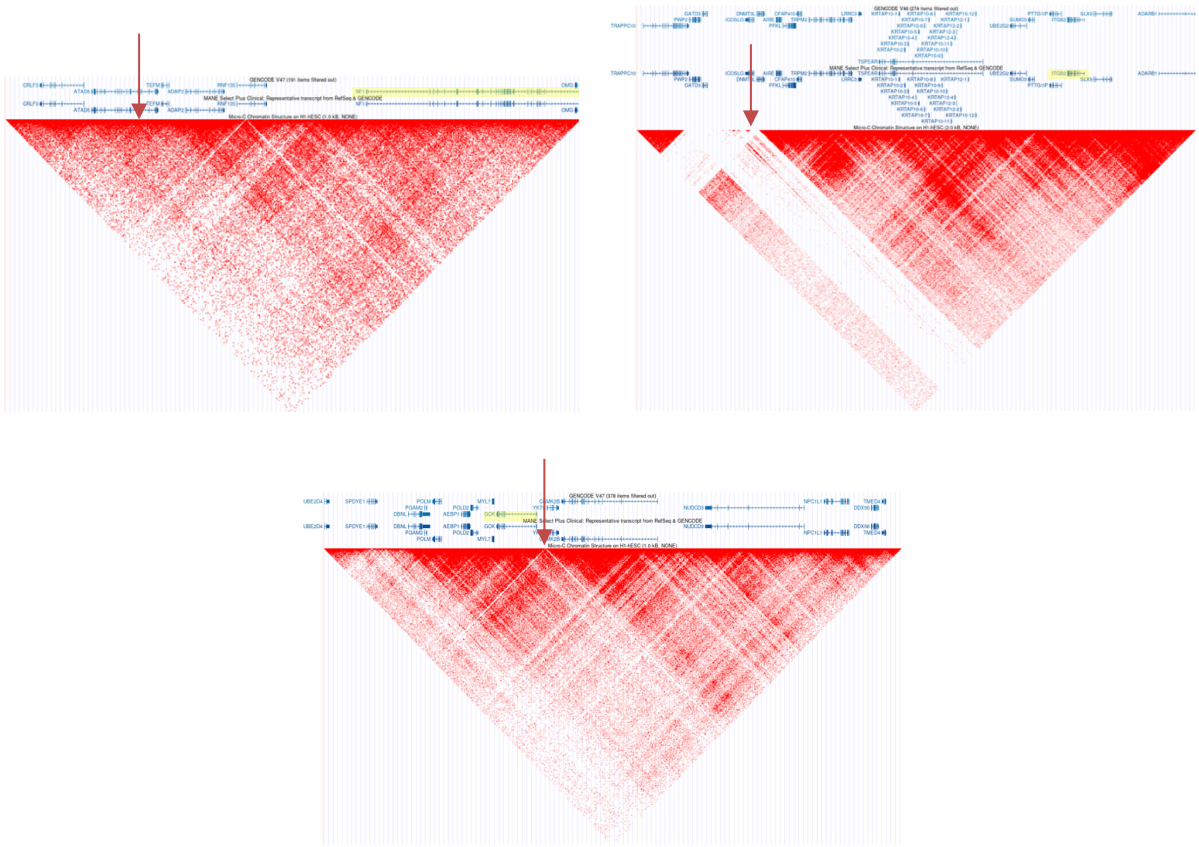


Figure 15: Hi-C maps of TADs where SNPs reside that uniquely identified genes for GeneCards and E-MAGO. *The highlighted gene is the identified gene, and the red arrow shows where the SNP is located. BC (top left), IBD (top right), T2D (bottom)*

The precision and recall for the three diseases were computed comparing GeneCards with E-MAGO and GREAT. Overall Figure 16 shows that E-MAGO has a higher precision than GREAT per disease. For BC and IBD, recall is also higher for E-MAGO. For T2D, recall is the same for both methods. This suggests that E-MAGO has the ability to capture more true positive genes in its predicted gene set than GREAT. For BC and IBD, E-MAGO was able to capture a higher fraction of true positive genes from its predicted genes than GREAT while T2D captured the same fraction for E-MAGO and GREAT.

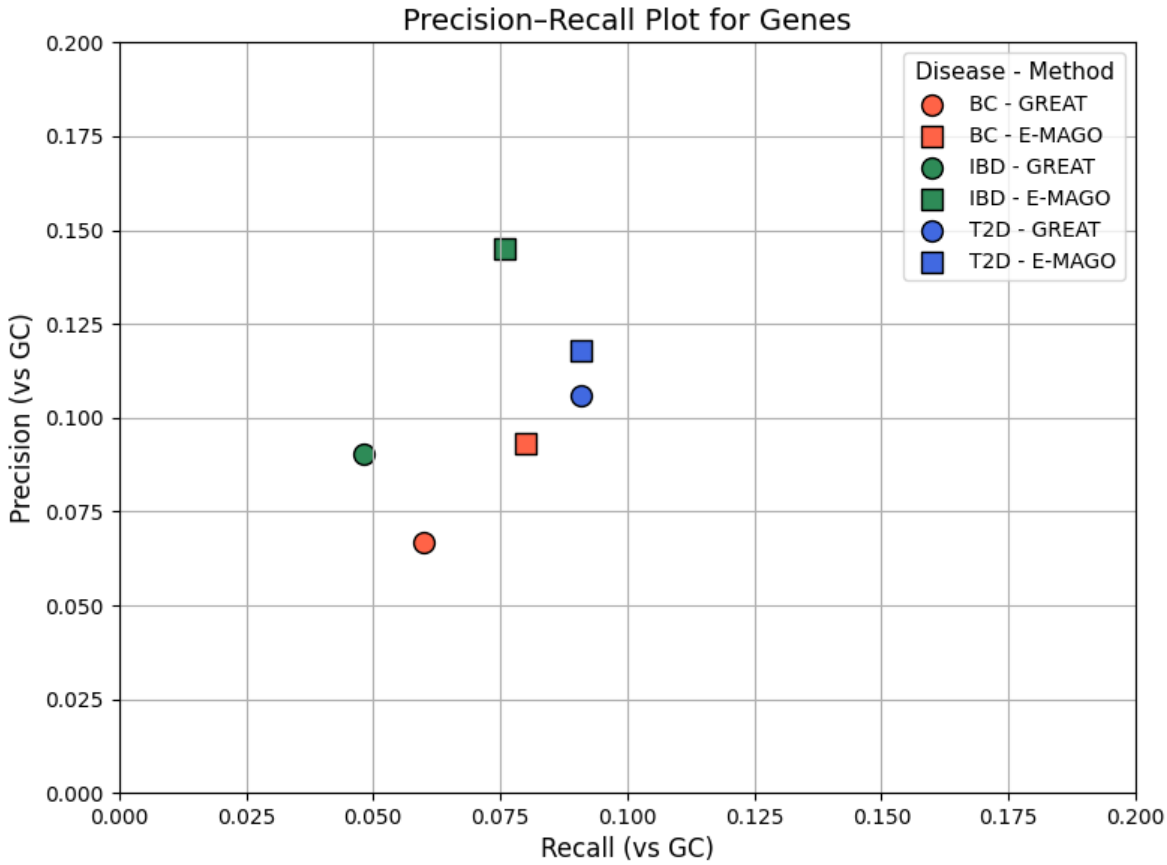


Figure 16: Scatterplot for the precision and recall of genes. *The X-axis is the recall, and the Y-axis is the precision.*

3.2.1.3 Overlap of GO terms

While GeneCards gene lists reflect what is already known about a disease, GWAS may generate new knowledge, e.g., novel implicated genes. To test if candidate genes annotated by E-MAGO contribute to the same biological processes associated with the disease-linked genes from GeneCards, GO enrichments were performed using g:GOSt. Following g:SCS correction, enriched GO terms with adjusted p-values below 0.05 were retained. The resulting terms were then compared across methods. Specifically, the overlap in GO terms was visualized in Figure 17. The results show that E-MAGO shares substantially more GO terms with GeneCards than GREAT does, suggesting that the genes identified by E-MAGO are more likely to be involved in disease-relevant biological processes compared to those found by GREAT.

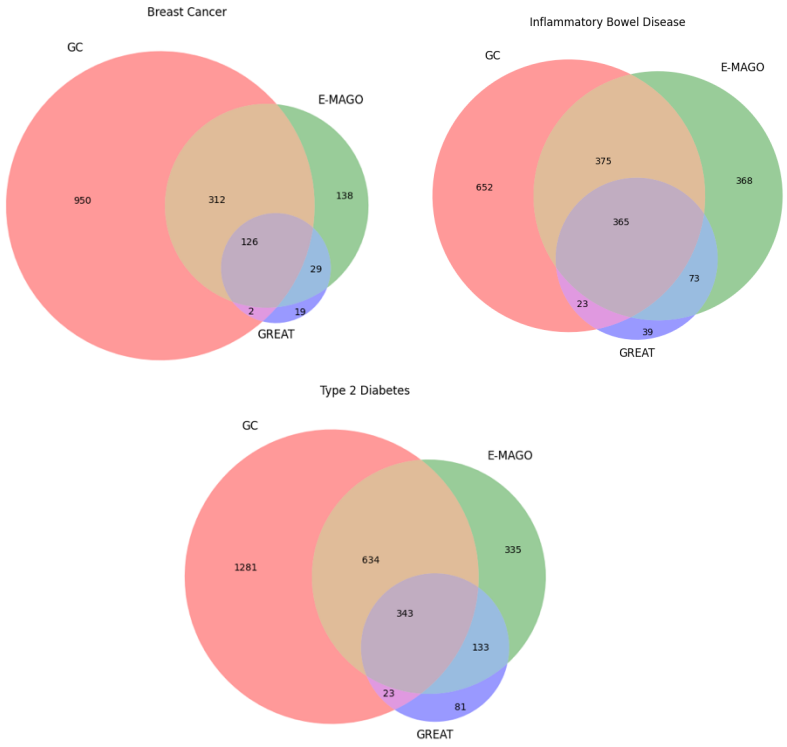


Figure 17: Venn diagrams visualizing the overlap in GO terms between GeneCards, E-MAGO and GREAT.

To better interpret these overlapping terms and assess their relevance to each disease, the GO terms were grouped into broader, high-level GO slim terms, as shown in Figure 18. This categorization facilitates easier comparison between the groups. E-MAGO shows more overlap with the most identified terms of GeneCards than GREAT. In BC, important procesess are cell proliferation and angiogenesis which are included in anatomical structure development, cell differentiation, immune system processes and programmed cell death. In IBD, terms like immune system processes, infammatory response and cell adhesion are important. In T2D, terms like metabolic processes, pancreatic development which is included in anatomical structure development, transport and signaling processes are important. For all these terms, E-MAGO showed a higher overlap with GeneCards. This suggests that E-MAGO has the potential to identify novel target genes acting on known disease-linked biological processes.

Figure 18: The GO terms found for GeneCards and the overlap of GO terms from E-MAGO and GREAT with GeneCards*. Illustrated with a stacked barplot with on the X-axis the descriptions of the GO slim terms and on the Y-axis the number of GO terms categorized into a certain GO slim term. Different colors show the amount of genes that share the same GO slim terms with GeneCards.*

The precision and recall for the GO terms of the three diseases were computed comparing GeneCards with E-MAGO and GREAT. Overall, E-MAGO has a higher precision than GREAT per disease. For BC, precision is equal for the two methods, while recall is higher for E-MAGO. This implies that E-MAGO finds a higher fraction of true positive terms in its predicted gene set. For IBD, precision is slightly higher while recall is much higher for E-MAGO. This implies that E-MAGO performs more accurately than GREAT for IBD. FOR T2D, recall is higher, but precision is lower for E-MAGO. This suggests that E-MAGO has the ability to find more true positive terms in its predicted set of terms, but it also captures more false positives than GREAT.

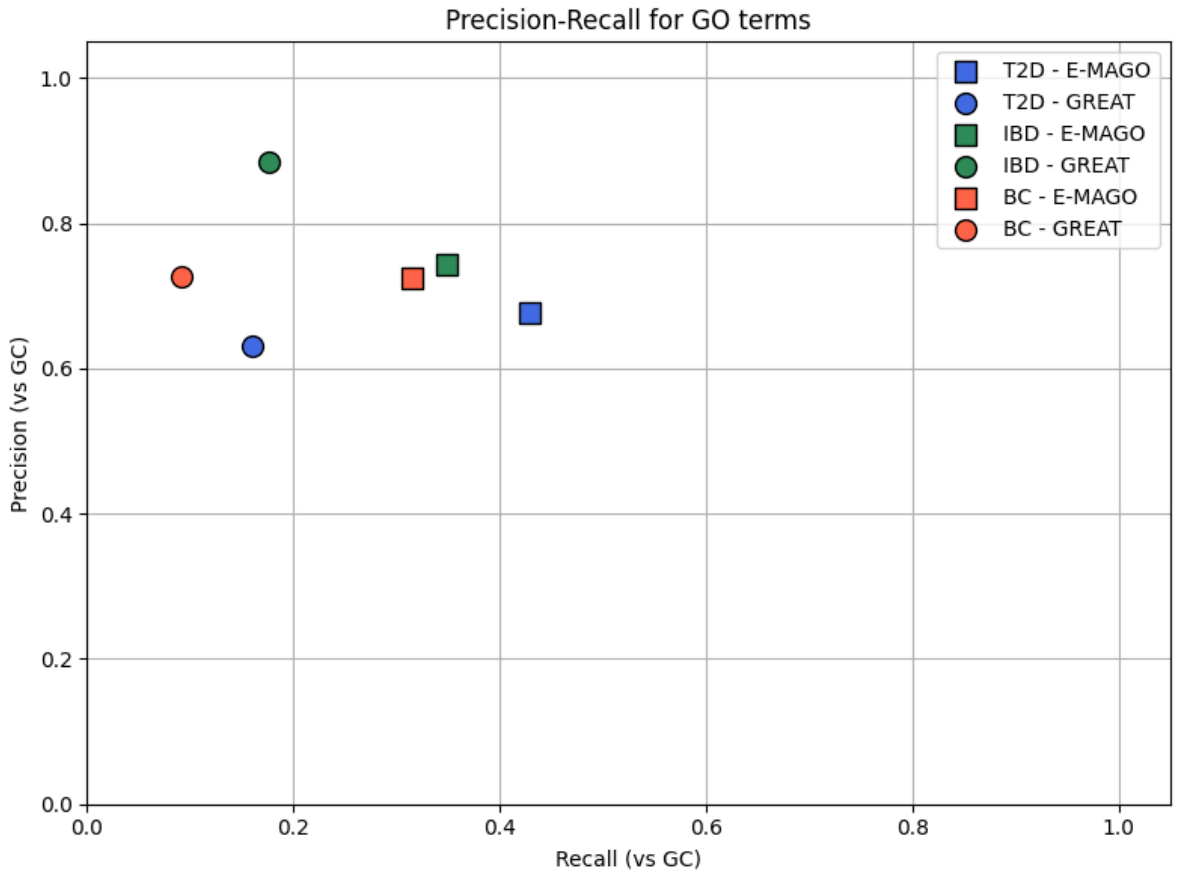


Figure 19: Scatterplot for the precision and recall of GO terms. *The X-axis is the recall, and the Y-axis is the precision.*

3.2.2 Validation with eQTL-based genes

The eQTL-GWAS co-localization strategy is considered the state-of-the-art technique for annotating genes to GWAS loci. By finding an eQTL and GWAS signal that share a causal variant, the expression of the gene according to the eQTL can serve as a connection between the gene and the disease and consequently identify the disease-linked gene for the GWAS locus. This co-localization strategy has shown to have high precision in linking GWAS loci to genes. Analogous to the use of GeneCards genes, the eQTL co-localization results should be considered ground truth while GREAT and E-MAGO are compared to it. However, it is currently not possible to colocalize every lead SNP with an eQTL. On the one side, not every lead SNP is a causal variant. Some of them are in LD with the causal variant meaning that they cannot possibly find an eQTL they co-localize with. On the other hand, eQTL co-localization is greatly underpowered, implying that co-localization results are scarce and retrieving a result for each causal leadSNP is impossible now.

Co-localization was done on BC, IBD and T2D using COLOC (Pullin & Wallace, 2024) with full summary statistics, lead SNPs to prioritize which SNPs are of importance and eQTL data from the adult GTEx project (Aguet et al., 2020a) filtered according to the relevant tissues for the studied disease. Table 2 demonstrates how 2, 4 and 3 genes were found for BC, IBD and T2D, respectively. These results show how greatly underpowered eQTL-GWAS co-localization is, making comparison with E-MAGO and GREAT less interpretable.

Table 2: Results of COLOC for BC, IBD and T2D.

disease	gene	variant_ID
BC	TOX3	chr16_52565276_C_T_b38
BC	L3MBTL3	chr6_130027974_C_T_b38
IBD	DDX11, OVOS2	chr12_31073901_T_A_b38
IBD	ERAP2, LNPEP	chr5_96916728_G_A_b38
T2D	MAN2C1, SNUPN	chr15_75458042_T_C_b38
T2D	CWF19L1	chr10_100152307_T_C_b38

As an alternative, eQTL data is obtained from the eQTL Catalogue, a comprehensive database that aggregates eQTLs from publicly available datasets. (Kerimov et al., 2021) By providing the lead SNP IDs and specifying the relevant tissue types, an API call retrieves all associated eQTLs along with their corresponding genes. For each lead SNP, the gene with the lowest p-value—indicating the most significant expression—is selected. This is not co-localization, but it is a look-up to find all genes an eQTL affects.

Important to note, in this experiment not all lead SNPs of a GWAS study are used. Only lead SNPs that are more than 50 kb away from their closest gene are used. Most regulatory elements are situated within 15 kb of the transcription start site of the gene resulting in 90% of the SNPs that affect the gene's expression situated within 15 kb. (Pickrell et al., 2010) This implies that every lead SNP in close proximity (< 50 kb) to a gene will overlap with an eQTL which will not guarantee co-localization, but it could also be because of LD. By filtering for lead SNPs with their closest gene more than 50 kb away, the focus is shifted to genes that are influenced by long-range regulatory elements. Lead SNPs with overlapping GWAS and eQTL signals will have a higher likelihood of being affected by distal regulatory effects and not by correlation with nearby gene expression signals due to LD.

This experiment uses the same GWASs for BC, IBD and T2D that were used in the first experiment. The filtering step results in 80, 48 and 94 lead SNPs being left, respectively. To compensate for this reduction, more GWASs were added for comparison namely, schizophrenia and Alzheimer's disease These diseases had 222 and 186 lead SNPs and after filtering they had 144 and 87 lead SNPs left which were all successfully annotated. This broader set of GWASs enhances the robustness of further analyses by expanding the pool of trait-associated variants that can be evaluated for potential long-range regulatory effects.

3.2.2.1 Overlap of genes

The overlap in genes between GREAT, E-MAGO and eQTL results for every disease were calculated. GREAT finds more overlapping genes with the eQTLs than E-MAGO over all diseases. Using this eQTL data, it suggests that lead SNPs are generally more linked to the genes in closest proximity and thus affect the expression of these closest genes.

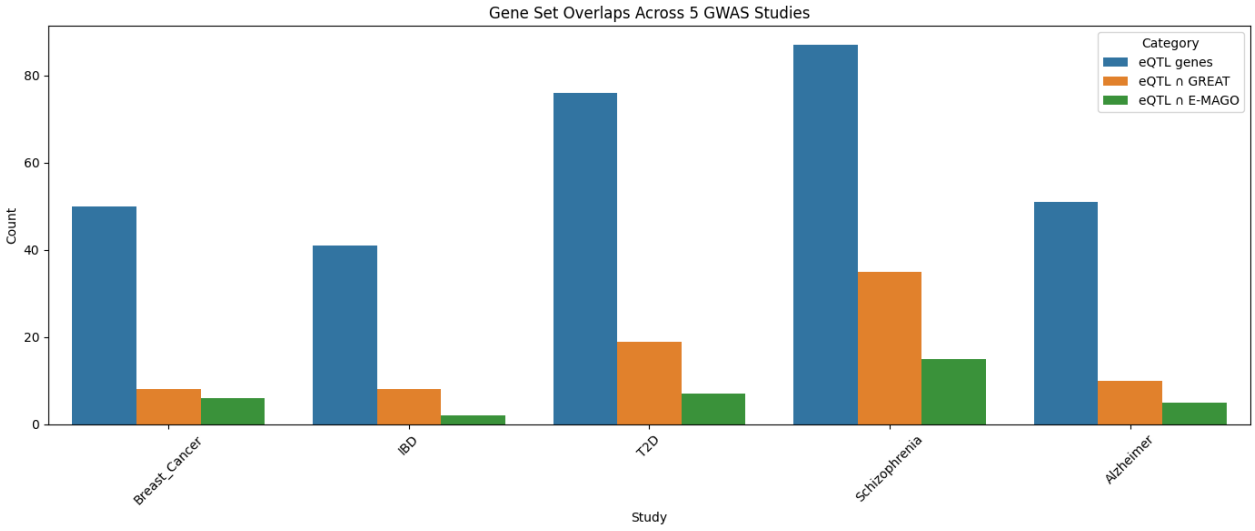


Figure 20: The comparison E-MAGO and GREAT genes with eQTL genes. *Illustrated in a grouped barplot with on the X-axis the diseases and on the Y-axis the number of genes.*

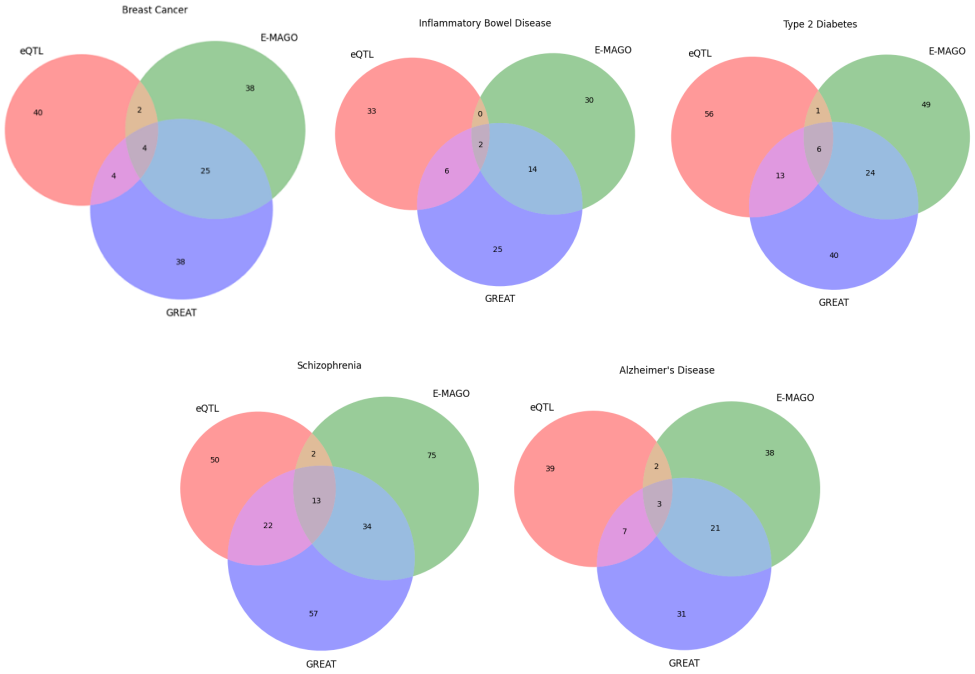


Figure 22: Venn diagrams visualizing the overlap in genes for eQTL overlap, E-MAGO and GREAT.

Precision and recall were calculated for E-MAGO and GREAT with the containing all the true positives. In every disease, GREAT has a higher precision and recall than E-MAGO. This implies that the genes predicted by GREAT are more likely to be truly disease-related and that GREAT finds more of the actual disease genes according to the eQTL data.

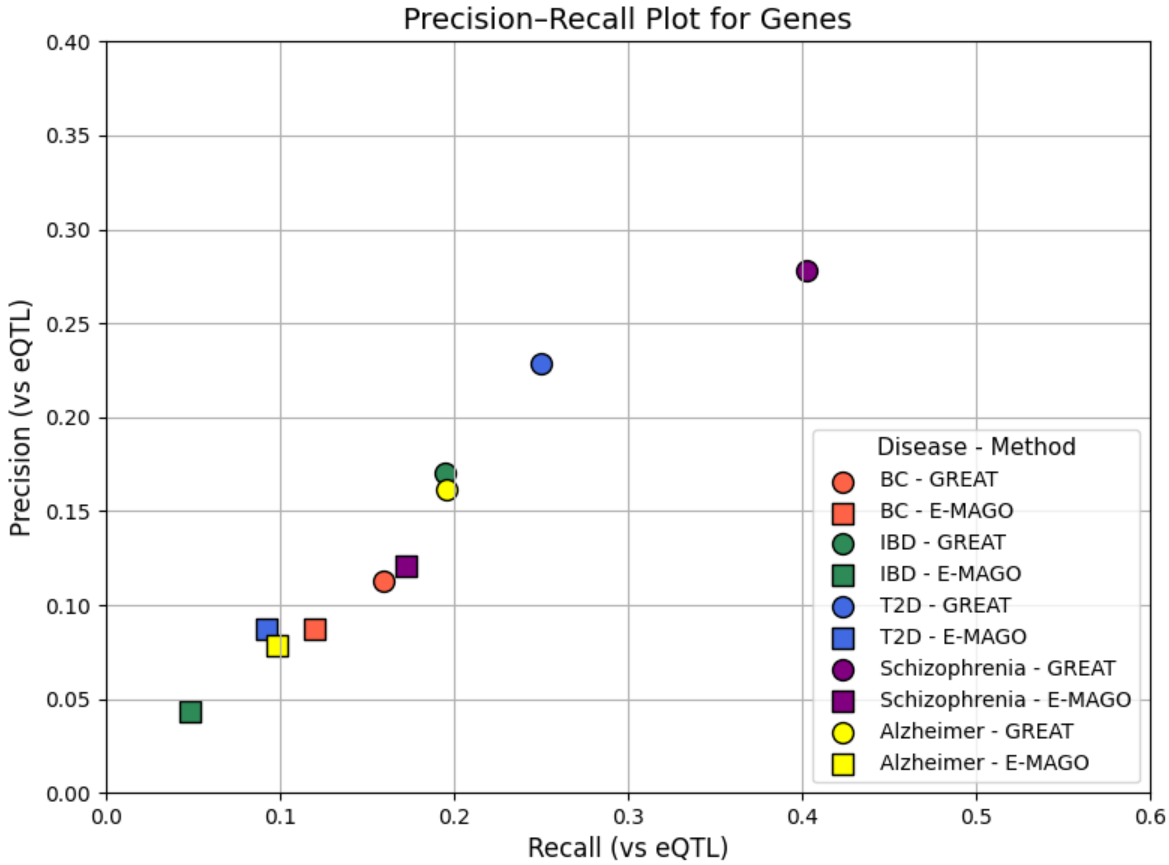


Figure 23: Scatterplot for the precision and recall of genes. *The X-axis is the recall, and the Y-axis is the precision.*

4 Discussion

This thesis presents the development of E-MAGO, a novel method for annotating genes to GWAS loci by incorporating biological context through GO enrichment. The method aims to improve gene prioritization precision compared to existing tools by iteratively refining gene relevance using an EM algorithm. E-MAGO was evaluated on multiple GWAS datasets from the GWAS Catalog and benchmarked against the widely used tool, GREAT, using disease-associated genes from GeneCards and eQTL-based genes as reference standards. E-MAGO demonstrated greater alignment with GeneCards genes than GREAT, suggesting improved identification of biologically relevant targets. In contrast, when eQTLs were used as ground truth, GREAT outperformed E-MAGO. This section discusses the results obtained during this thesis.

4.1 E-MAGO enhances target gene annotation beyond proximity

4.1.1 E-MAGO's biological relevance results in closer alignment with GeneCards

The genes identified by E-MAGO showed greater overlap with GeneCards genes compared to those found using GREAT across all studies, except for T2D, where the overlap was equal. Notably, genes uniquely detected by E-MAGO and GeneCards were located in gene-dense TADs. For lead SNPs within those TADs, GREAT failed to identify the disease-associated genes from GeneCards. This is because GREAT always assigns SNPs to the nearest gene, without considering alternative candidates. In contrast, E-MAGO begins with the closest gene but iteratively shifts to the gene that is biologically more relevant. Genes identified solely by GeneCards and GREAT may reflect cases where GREAT coincidentally assigned a disease-linked gene simply because it was nearby. However, enhancers can be located near the promoter of one gene but interact with another gene that is linearly more distant, due to the 3D organization of the genome bringing enhancer regions into contact with different promoters. As a result, a SNP may appear linked to a nearby gene by chance and overlap with a gene listed in GeneCards, while E-MAGO identifies a more likely disease-relevant target gene for that SNP. Genes detected by all three methods—GeneCards, E-MAGO, and GREAT—are often both the nearest and the biologically relevant ones. These SNPs most likely affect promoter regions, nearby enhancer elements, or TSSs, or may reside within the gene itself, directly influencing its translation. Interestingly, the top-scoring genes in GeneCards were not identified by either E-MAGO or GREAT, as they were in TADs without any lead SNPs. Consequently, these genes were not included in the search space.

The GO terms found for the different gene sets showed that E-MAGO had more overlapping terms with GeneCards than GREAT for all three studies. The result for BC and IBD makes sense, because having more common genes would mean having more common GO terms enriched in the gene sets. For T2D, more GO term overlap suggests that the genes found with E-MAGO were biologically more relevant than the genes found with GREAT. The strong overlap between GeneCards and E-MAGO demonstrates that E-MAGO is able to find more disease-relevant pathways. Notably, GREAT always had one third of the number of significant terms that E-MAGO had which shows that E-MAGO had a wider variety of biological processes involved in its gene set. This variety resulted in E-MAGO being able to capture more mechanisms relevant for the disease that GREAT had missed and to denoise the resulting genes from biologicallly irrelevant genes. This suggests that focusing on biological context is a promising approach to find disease-relevant genes for GWAS results.

4.1.2 Distal regulation insights through eQTL-based validation

The genes identified by E-MAGO showed less overlap with the genes found through eQTLs than GREAT. Only SNPs with their closest gene more than 50 kb away were used to make sure eQTLs residing in promotor regions, TSSs or genes were excluded focusing only on SNPs that affect enhancer regions located further away from the target gene. This distance helps to decrease the confounding effects of LD, which often makes it difficult to distinguish whether an eQTL signal originates from a promoter or from a more distal regulatory element. For genes uniquely overlapping with E-MAGO and eQTL look-up, the same interpretation as for the first experiment can be concluded, namely in gene-dense TADs GREAT fails to capture target genes expressed further away from the SNP. Genes uniquely identified by GREAT and eQTL look-up could imply that the SNP is located in the promotor region. However, these genes also suggest that when eQTLs are not located in promotor regions or TSSs of nearby genes they mostly affect the genes that are still closest to them. This could imply that enhancers are still close to their target promotor and that other more distal enhancer regions were not found. It was observed that over 90% of eQTLs are located closer than 15 kb from the TSS of the gene they impact. (Pickrell et al., 2010) Then, it is expected that also the whole region around the gene will show the same eQTL signal because of LD. Though, there are still many eQTLs that did not overlap with E-MAGO or GREAT. From the perspective of GREAT, this means that these eQTLs do lie in distal regulatory elements. From the perspective of E-MAGO, this implies that biological relevant genes found through GO enrichment might still have missed the true target genes because of incomplete GO annotations for less well-studied genes causing bias or that the search space within the TADs was too small.

It is important to consider that no real co-localization was done to find the eQTLs that overlap with the lead SNPs. In other words, overlap between an eQTL and GWAS signal was considered without testing if both signals had the same underlying causal SNP. Instead, a p-value look-up was done via the eQTL Catalogue making the eQTL genes more prone to LD confounding. Also, the eQTL Catalogue only contains cis-eQTLs which means that eQTLs are no further than 1 Mb away from their target gene. E-MAGO uses TADs that extend over several megabases far. In those TADs, E-MAGO spans a wider search space compared to the eQTL Catalogue, a limitation that could have negatively impacted E-MAGO's performance scores, even if E-MAGO had identified the true target gene.

4.2 Limitations of eQTLs for validating E-MAGO

The use of cis-eQTLs limits the analysis to genes located near the genetic variant which limits the value of the validation. Although, it has been observed that the affected gene typically lies within the same TAD which provides a biological meaningful boundary for regulatory interactions. (Aguet et al., 2020b) E-MAGO could have the potential to identify the same target genes, since its search space is also confined to TAD boundaries. However, this represents only part of the regulatory landscape. Notably, cis-eQTLs account for only about 30% of gene expression heritability. (Umans et al., 2021), indicating that they miss many distal enhancer–promoter interactions that play critical roles in gene regulation and biological pathways. Trans-eQTLs have the potential to capture these long-range interactions, making them especially relevant for uncovering the true target genes, and thus providing a more comprehensive comparison to E-MAGO. However, trans-eQTLs are much less studied than cis-eQTLs, primarily because their effect sizes are typically smaller and harder to detect. Interestingly, many trans-eQTLs are influenced by cis-mediated effects, where a variant affects a nearby gene (a cis-eQTL), which in turn influences a distant gene. In fact, a recent study found that 77% of trans-eQTLs (from the 31.6%

that were linked to a gene) appeared to act indirectly through cis-regulated genes. (Aguet et al., 2020b) Thus, the use of cis-mediated trans-eQTLs could further complicate finding the target gene as it might not point to its true target gene but an intermediate one. Another study investigated independent trans-eQTLs to determine their role in complex traits and diseases. It found that only 4.8% of all detected eQTLs were trans, and even fewer were truly independent. As a result, only a small fraction of complex traits or diseases could be attributed to trans-eQTLs. However, some significant associations were observed—for example, with height and rheumatoid arthritis—underscoring the importance of further research into trans-eQTLs to better understand distal regulatory mechanisms which would enhance validation of E-MAGO with eQTLs. (Yap et al., 2018)

An important characteristic of eQTLs is their spatiotemporal specificity—their effects can vary depending on the tissue, cell type, and developmental stage. Therefore, careful consideration must be given to selecting the appropriate context when collecting eQTLs for comparison, especially in relation to the trait or disease being studied. For example, eQTL studies on craniofacial traits and diseases require craniofacial tissue from embryonic developmental stages which cannot be accessed very easily due to ethical and technical constraints. (Naqvi et al., 2022) In such cases, adult tissues may be less informative, as gene expression levels often remain stable and do not reflect the dynamic changes occurring in early life. (Umans et al., 2021) Such traits and diseases cannot accurately map eQTLs to genes meaning that these are difficult to use for validating E-MAGO. In general, variants located in coding regions tend to have consistent effects across different tissue types, whereas variants in non-coding regions—such as promoters, enhancers, and TSSs—are more likely to exhibit tissue-specific effects. Accurately identifying the relevant tissue types for a given study requires prior biological knowledge, and even with such knowledge, it is still possible to overlook the true tissue of action, especially in the case of less well-studied diseases. Beyond tissue type, differences in cell type composition can also lead to distinct gene expression patterns. The use of bulk tissue eQTLs introduces a mixture of signals from various cell types, potentially blurring cell type-specific effects. However, tissue types that share similar cell type architectures tend to exhibit similar eQTL profiles, suggesting that shared cellular composition underlies shared regulatory variation. (Aguet et al., 2020b) Carefully choosing tissue and cell types is critical for accurate validation of E-MAGO using eQTLs.

Validation using genes identified through eQTL co-localization is challenging due to the limited statistical power of available eQTL data in combination with the low genomic resolution imposed by LD structure. As mentioned above**,** not all relevant tissue types are accessible depending on the spatiotemporal contexts. Moreover, trans-eQTLs account for a substantial portion of gene expression heritability, but their small effect sizes and the need for very large sample sizes introduce a heavy multiple testing burden, making it difficult to detect statistically significant associations. (Zhou & Cai, 2021) Attempts to use tools for true co-localization identified only one or a few genes with very high precision, but these results were too limited to allow for meaningful or interpretable comparisons.

4.3 TADs as stable and biologically informed framework

The TADs used in E-MAGO were sourced from TADKB, which integrates Hi-C data and combines outputs from two clustering algorithms. These domains define gene search spaces that range from a few genes to over 50 genes and span hundreds of kilobases to several megabases, allowing the detection of long-range regulatory interactions. This makes TADs a biologically meaningful framework for limiting the candidate gene space in functional analyses. However, in the first experiment, it was

noted that several high-scoring disease-linked genes did not fall within the same TADs as the lead SNPs. A possible explanation is that the TADs used were not broad enough to capture distal regulatory interactions—certain lead SNPs may indeed regulate these genes, but their exclusion from the TAD-defined search space led to them being missed. This reflects the limitations of clustering methods on Hi-C data, which struggle to detect higher-order TAD boundaries that span broader genomic regions and could better capture these long-range regulatory effects. (Rocha et al., 2015)

Despite some variability across cell types and developmental stages, TAD boundaries are relatively well conserved. TAD boundary stability was investigated across 37 diverse cell types and found that the majority of TAD boundaries are shared across these cell types. Notably, boundaries stable across many cell types are significantly enriched for functionally and evolutionary constrained regions like CTCF binding and core genes. These stable boundaries overlapped with 5,420 conserved base pairs on average, demonstrating their evolutionary constraint. (McArthur & Capra, 2021) This makes E-MAGO substantially less dependent on cell type-specific data compared to eQTL-based gene mapping strategies, which require expression data from relevant tissues. Moreover, tailoring the TAD map to match the disease-relevant cell type may offer modest performance gains, but the overall conservation of TAD boundaries ensures that E-MAGO remains broadly applicable. Fortunately, TAD maps are available for a wide variety of cell types through resources such as TADKB.

Some results, such as genes uniquely identified by both GREAT and GeneCards, suggest that GREAT may have coincidentally identified a lead SNP near a disease-linked gene. GREAT is limited to assigning SNPs to the nearest gene and does not account for biological relevance, which is why it is outperformed by E-MAGO in identifying disease-associated genes. Proximity alone does not guarantee regulatory influence. It is possible that the SNP lies within an enhancer region close to one gene while actually regulating a more distal target. In fact, over 40% of enhancers bypass their nearest promoters and instead regulate distant genes, highlighting that linear distance is often a poor predictor of gene regulation. (Li et al., 2012)

Some other findings are that genes that reside alone within a TAD tend to be more conserved and stable than those sharing a TAD with other genes. This suggests that solitary genes may have critical functions and are isolated within a TAD to protect their expression from disruptive elements. These are the genes that will be captured by both GREAT and E-MAGO if it is the only gene near the lead SNP and the only gene inside the TAD. (Long et al., 2022) Finally, it should also be considered that one enhancer can regulate multiple genes, and a single gene can be regulated by multiple enhancers. (Schoenfelder & Fraser, 2019) This implies that different lead SNPs inside the same TAD—whether in linkage disequilibrium or affecting separate enhancers—can ultimately converge on the same gene. When two lead SNPs in a TAD link to the same target gene, E-MAGO will always choose for the same gene in the TAD (the highest scoring gene). In contrast, GREAT may assign them to different genes, as it always selects the nearest gene, which is not necessarily the same for both SNPs.

4.4 Influence of incomplete data

E-MAGO relies on GO data and a set of protein-coding genes obtained from the UCSC Genome Browser to perform GOEA. This means that the quality and accuracy of the results are entirely dependent on the available annotations and gene sets provided to the method. However, it is important to note that this information is inherently incomplete. Because genomic knowledge is constantly evolving, many functional elements and regulatory genes remain functionally invalidated or less well-validated which

implies that the GO annotations remain incomplete. As a result, the actual target genes of lead SNPs may be less likely to appear through GO enrichment and consequently missed—limiting the potential of E-MAGO to identify them.

The GO database itself has undergone significant growth. Between 2004 and 2015, the number of GO terms are multiplied by a factor of 2.5, and the number of annotations expanded by a factor of 6.3, leading to many new gene-term associations. (Tomczak et al., 2018) As of today, the GO database contains over 8.6 million annotations, but only around one million of these are backed by experimental evidence, highlighting that most annotations are based on inference and carry lower confidence. In 2018, there were still only about 700 000 experimentally validated annotations, underlining the ongoing gap in high-quality functional data. (*Gene Ontology Resource*, n.d.) Although the GO resource continues to improve, its limitations must be considered when interpreting enrichment results in tools like E-MAGO.

The protein-coding gene set used by E-MAGO comes from the GENCODE project, integrated through the UCSC Genome Browser. GENCODE currently annotates 78 686 genes, of which 19 435 are protein-coding. Over the past 11 years, GENCODE has grown by nearly 20 000 genes, reflecting both better genome annotation techniques and new gene discoveries. (Mudge et al., 2025) This underscores the need to frequently update gene databases and serves as a reminder that many genes—and thus potential disease-relevant targets—may not yet be annotated or even known. Consequently, E-MAGO may overlook key genes simply because they are not yet part of the search space.

4.5 Suggestions for improving E-MAGO

E-MAGO sets the foundation for developing a high-precision gene annotation method that functions through an EM algorithm and prioritizes its genes based on the biological context given through GO enrichment. E-MAGO is not without shortcomings, but it does already describe the general idea of the strategy. Several improvements are proposed to fine-tune E-MAGO's precision.

First, genes are initialized based on linear proximity to the lead SNP, with all candidate genes (i.e., all coding genes connected to the lead SNPs, not only the initial genes) starting with a score of 0. This assumes that every gene has an equal chance of being the true target, though in reality, this is unlikely. Probabilistic modelling could improve initialization by taking into account prior knowledge. Depending on the phenotype being investigated, prior information about relevant biological pathways could be used to prioritize GO terms associated with those processes. Genes linked to these user-prioritized GO terms could then be assigned higher scores accordingly.

Second, E-MAGO is biased for well-known genes which have a high number of GO annotations. These genes are more likely to be scored higher as they appear more frequently. To solve this problem, the number of GO annotations per gene could be taken into account. Genes with less annotations which are less well-known genes can receive more weight when they appear after GO enrichment.

Third, as mentioned above, a major proportion of GO annotations are not experimentally validated. These lower-confidence annotations could reduce the precision of the method. To address this, GO annotations supported by experimental evidence could be weighted more heavily than those inferred computationally.

Fourth, the TAD data used may not always define gene search spaces robustly enough to include true target genes. Rather than applying hard TAD boundaries that sharply exclude genes outside the domain, E-MAGO could be extended to incorporate soft TAD boundaries based on contact frequencies. In this framework, genes located just outside a TAD boundary would not be excluded outright, but instead assigned a lower probability of being the target. This approach would reflect the biological uncertainty of domain boundaries and allow regulatory interactions near TAD edges to be captured. By integrating boundary flexibility into its scoring system, E-MAGO could become more sensitive to distal, yet biologically plausible, gene targets.

5 Conclusion

This thesis has developed a novel method for annotating genes to GWAS loci called E-MAGO. By comparing it to a distance-based approach, the new strategy's performance was validated. The first comparison was done with GeneCards genes as ground truth which showed that E-MAGO performs better than distance-based annotation. The second comparison was done with eQTL-based annotation genes as ground truth. This showed the opposite namely, that E-MAGO performed worse than a distance-based approach.

The first comparison highlighted that in gene-dense regions where the target gene is located further away, E-MAGO could find target genes that the distance-based method could not. Demonstrating that the 3D architecture of the genome is of high importance to link SNPs on enhancer regions located far away from the target gene. It also showed that high scoring disease-linked genes were not identified by any method which was because none of the GWAS loci was close enough to the genes to be captured by the distance-based approach or to be included in the search space of E-MAGO.

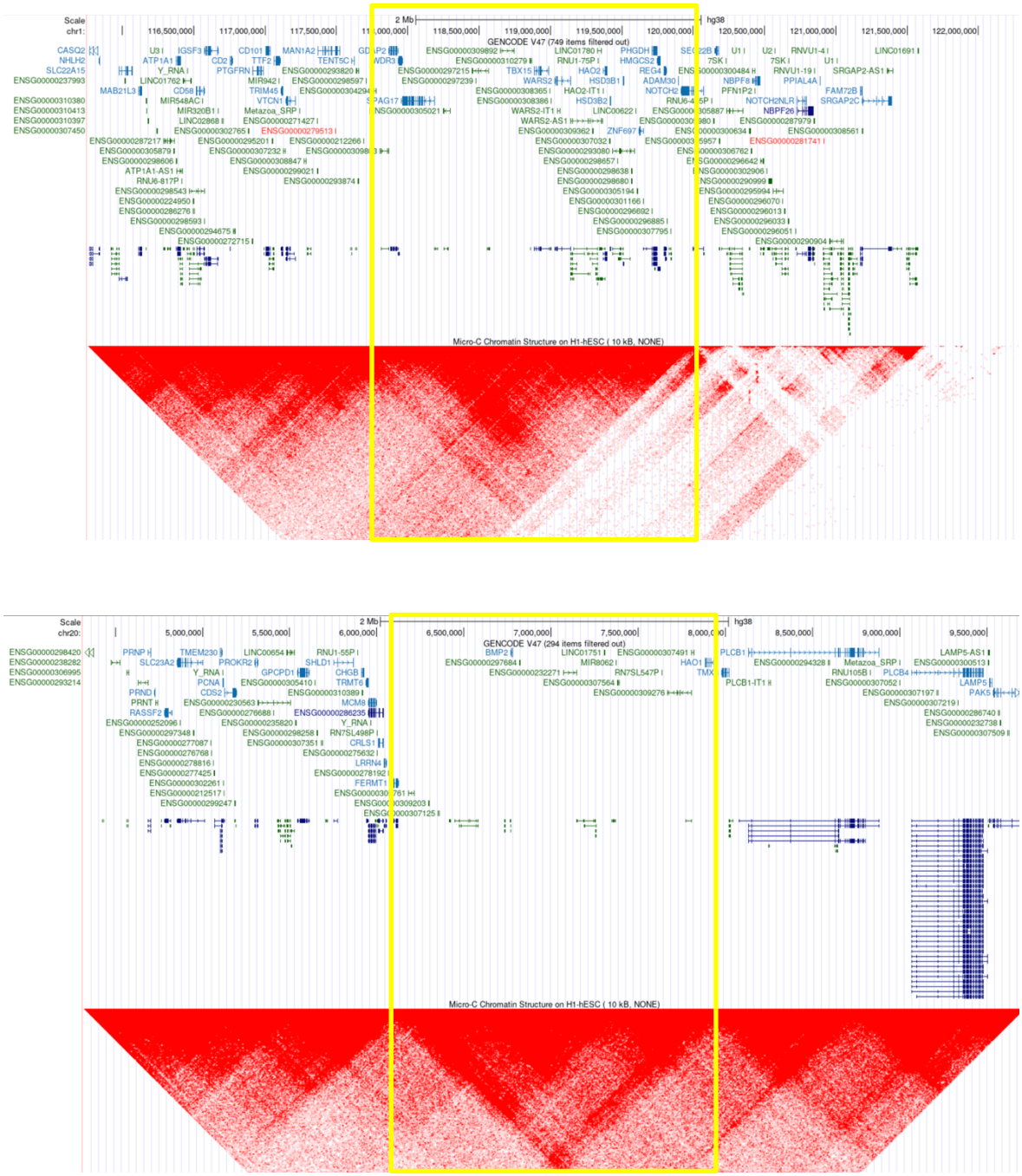
The second comparison highlighted the need for using genes found through eQTL-GWAS co-localization as ground truth. The second experiment did not use co-localization results but just a look-up of cis-eQTLs that overlap with the lead SNPs. All SNPs with genes closer than 50 kb were filtered out to make sure that eQTLs in LD were excluded. This was an attempt to make the eQTL genes more reliable, but it did not yield any clear validation. This emphasizes that E-MAGO's performance is limited when benchmarked against cis-eQTL genes that did not result from co-localization analysis.

Annotating target genes to their GWAS loci using E-MAGO is a promising approach, as it incorporates biological context and leverages TADs to account for the 3D organization of the genome. However, the method cannot fully realize its potential due to several limiting factors: the use of incomplete data, such as the input gene sets and GO annotations; the clustering methods to find realistic TADs, which is crucial for accurately capturing the regulatory architecture relevant to the studied phenotype; and the lack of trans-eQTLs and eQTL-GWAS co-localization analyses, which are essential for robust benchmarking.

In future work, improvements such as integrating phenotype-specific prior knowledge, applying weighted GO evidence and using dynamic TAD boundaries could enhance E-MAGO's precision and robustness.

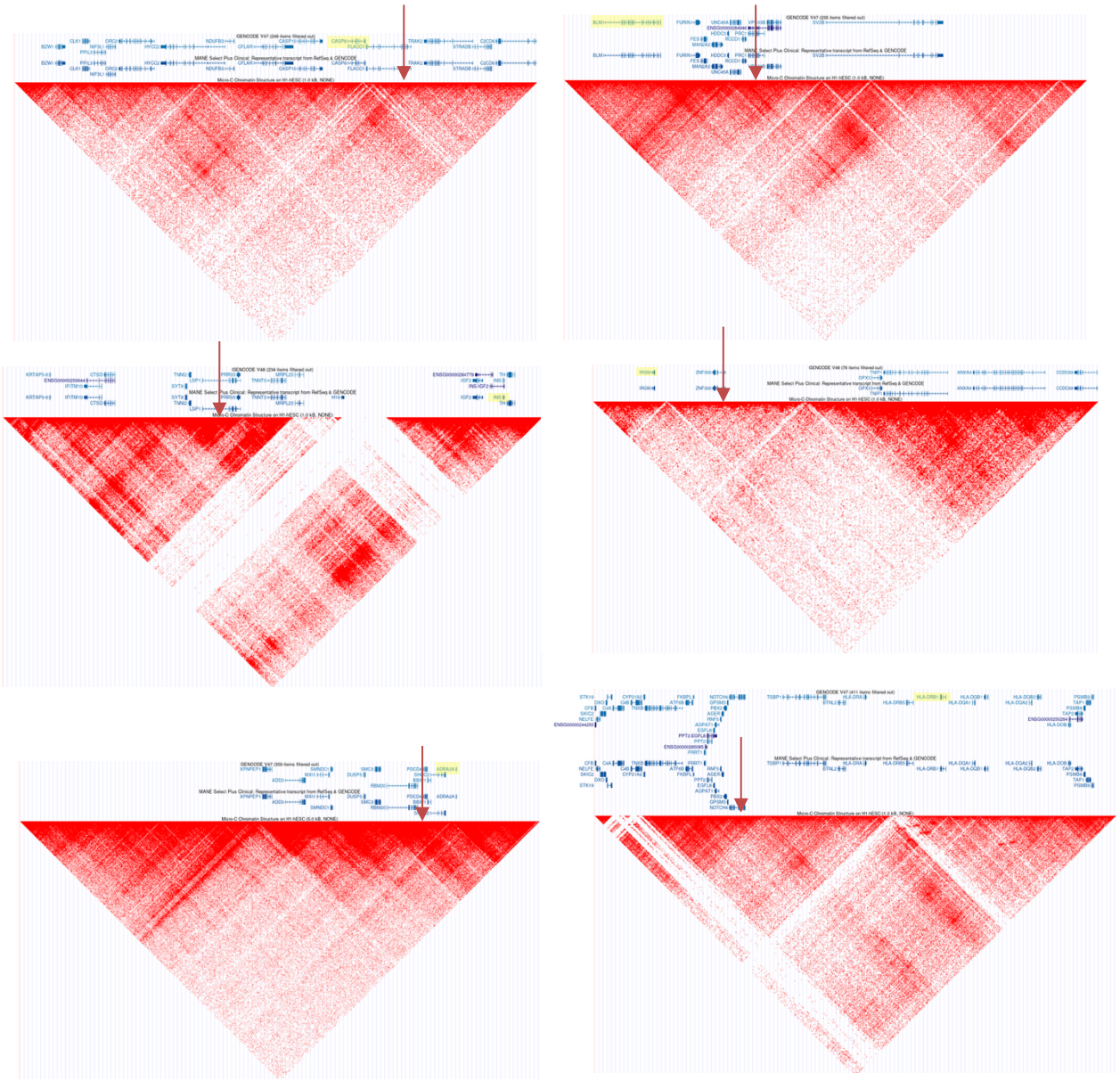
6 Supplementary data

6.1 Manual validation of TADKB's TAD boundaries



Supplementary figure 1: Demonstration of TADKB's TAD boundaries in the UCSC genome browser. *The yellow frames show where TADKB determined their boundaries.*

6.2 Gene-dense regions



Supplementary figure 2: Hi-C maps of TADs where SNPs reside that uniquely identified genes for GeneCards and E-MAGO. *The highlighted gene is the identified gene, and the red arrow shows where the SNP is located. BC (top), IBD (middle), T2D (bottom)*

7 References

Adewuyi, E. O., O'Brien, E. K., Nyholt, D. R., Porter, T., & Laws, S. M. (2022). A large-scale genome-wide cross-trait analysis reveals shared genetic architecture between Alzheimer's disease and gastrointestinal tract disorders. *Communications Biology*, *5*(1). https://doi.org/10.1038/S42003-022-03607-2,

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., … Volpi, S. (2020a). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/SCIENCE.AAZ1776/SUPPL_FILE/AAZ1776_TABLESS10-S16.XLSX

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Flynn, E. D., Parsana, P., Fresard, L., Gamazon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., … Volpi, S. (2020b). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (New York, N.Y.)*, *369*(6509), 1318. https://doi.org/10.1126/SCIENCE.AAZ1776

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, *40*(3), 340–345. https://doi.org/10.1038/ng.78

Benoit, K. (2011). *Linear Regression Models with Logarithmic Transformations*.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. In *Cell* (Vol. 169, Issue 7, pp. 1177–1186). Cell Press. https://doi.org/10.1016/j.cell.2017.05.038

Brodie, A., Azaria, J. R., & Ofran, Y. (2016). How far from the SNP may the causative genes be? *Nucleic Acids Research*, *44*(13), 6046–6054. https://doi.org/10.1093/nar/gkw500

Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., … Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/NAR/GKY1055,

Central Dogma. (n.d.). Retrieved May 25, 2025, from https://www.genome.gov/genetics-glossary/Central-Dogma

Chen, Z., Boehnke, M., Wen, X., & Mukherjee, B. (2021). Revisiting the genome-wide significance threshold for common variant GWAS. *G3: Genes, Genomes, Genetics*, *11*(2). https://doi.org/10.1093/g3journal/jkaa056

Crouch, D. J. M., & Bodmer, W. F. (2020). *Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants*. *117*(32), 18924–18933. https://doi.org/10.1073/pnas.2005634117/-/DCSupplemental

De Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S. G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/NG.3760;SUBJMETA=1503,205,208,2138,257,631,692,699;KWRD=GENOME-WIDE+ASSOCIATION+STUDIES,INFLAMMATORY+BOWEL+DISEASE

Dehghan, A. (2018). Genome-wide association studies. In *Methods in Molecular Biology* (Vol. 1793, pp. 37–49). Humana Press Inc. https://doi.org/10.1007/978-1-4939-7868-7_4

Dierckx, T., & Moreau, Y. (n.d.). *Bayesian Modeling for Biological Data Analysis Course Notes*.

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, *26*(8), 897–899. https://doi.org/10.1038/NBT1406;KWRD=LIFE+SCIENCES

Expectation-maximization: theory and intuition - Matthew N. Bernstein. (n.d.). Retrieved May 26, 2025, from https://mbernste.github.io/posts/em/

Fraser, J., Williamson, I., Bickmore, W. A., & Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*, *79*(3), 347–372. https://doi.org/10.1128/mmbr.00006-15

Gallagher, M. D., & Chen-Plotkin, A. S. (2018). The Post-GWAS Era: From Association to Function. In *American Journal of Human Genetics* (Vol. 102, Issue 5, pp. 717–730). Cell Press. https://doi.org/10.1016/j.ajhg.2018.04.002

Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K. K., Nasser, J., Jagadeesh, K. A., Weiner, D. J., Shi, H., Fulco, C. P., O'Connor, L. J., Pasaniuc, B., Engreitz, J. M., & Price, A. L. (2022). Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, *54*(6), 827–836. https://doi.org/10.1038/s41588-022-01087-y

Gene Ontology Resource. (n.d.). Retrieved May 25, 2025, from https://geneontology.org/stats.html

Génin, E. (2020). Missing heritability of complex diseases: case solved? In *Human Genetics* (Vol. 139, Issue 1, pp. 103–113). Springer. https://doi.org/10.1007/s00439-019-02034-4

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, *10*(5). https://doi.org/10.1371/JOURNAL.PGEN.1004383,

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946–1978. https://doi.org/10.1002/SIM.6082;PAGE:STRING:ARTICLE/CHAPTER

Gregor Mendel and the Principles of Inheritance | Learn Science at Scitable. (n.d.). Retrieved May 26, 2025, from https://www.nature.com/scitable/topicpage/gregor-mendel-and-the-principles-of-inheritance-593/

Gruber, S. B. (2007). Population stratification in epidemiologic studies of founder populations. *Cancer Biomarkers*, *3*(3), 123–128. https://doi.org/10.3233/CBM-2007-3302,

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., & Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, *109*(2), 83–90. https://doi.org/10.1016/j.ygeno.2017.01.005

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews. Molecular Cell Biology*, *19*(10), 621. https://doi.org/10.1038/S41580-018-0028-8

Han, J., Zhang, Z., & Wang, K. (2018). 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics 2018 11:1*, *11*(1), 1–10. https://doi.org/10.1186/S13039-018-0368-2

Heinig, M. (2018). Using Gene Expression to Annotate Cardiovascular GWAS Loci. In *Frontiers in Cardiovascular Medicine* (Vol. 5). Frontiers Media S.A. https://doi.org/10.3389/fcvm.2018.00059

Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics*, *99*(6), 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003

Huerta-Chagoya, A., Schroeder, P., Mandla, R., Li, J., Morris, L., Vora, M., Alkanaq, A., Nagy, D., Szczerbinski, L., Madsen, J. G. S., Bonàs-Guarch, S., Mollandin, F., Cole, J. B., Porneala, B., Westerman, K., Li, J. H., Pollin, T. I., Florez, J. C., Gloyn, A. L., … Mercader, J. M. (2024). Rare variant analyses in 51,256 type 2 diabetes cases and 370,487 controls reveal the pathogenicity spectrum of monogenic diabetes genes. *Nature Genetics 2024 56:11*, *56*(11), 2370–2379. https://doi.org/10.1038/s41588-024-01947-9

Interactive Fisher's Exact Test. (n.d.). Retrieved May 26, 2025, from https://quantpsy.org/fisher/fisher.htm

Jo, B.-S., & Choi, S. S. (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics & Informatics*, *13*(4), 112. https://doi.org/10.5808/gi.2015.13.4.112

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., Weber, R. J., Haussler, D., & Kent, W. J. (2003). The UCSC Genome Browser Database. *Nucleic Acids Research*, *31*(1), 51–54. https://doi.org/10.1093/NAR/GKG129,

Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature Genetics 2021 53:9*, *53*(9), 1290–1299. https://doi.org/10.1038/s41588-021-00924-w

Kim, S., & Shendure, J. (2019). Mechanisms of Interplay between Transcription Factors and the 3D Genome. In *Molecular Cell* (Vol. 76, Issue 2, pp. 306–319). Cell Press. https://doi.org/10.1016/j.molcel.2019.08.010

Kingdom, R., & Wright, C. F. (2022). Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. In *Frontiers in Genetics* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fgene.2022.920390

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389. https://doi.org/10.1126/science.1109557

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Vesztrocy, A. W., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28948-z

Kockum, I., Huang, J., & Stridh, P. (2023). Overview of Genotyping Technologies and Methods. *Current Protocols*, *3*(4). https://doi.org/10.1002/cpz1.727

Komar, A. A. (Ed.). (2009). *Single Nucleotide Polymorphisms*. *578*. https://doi.org/10.1007/978-1-60327-411-1

Krietenstein, N., Abraham, S., Venev, S. V., Abdennur, N., Gibcus, J., Hsieh, T. H. S., Parsi, K. M., Yang, L., Maehr, R., Mirny, L. A., Dekker, J., & Rando, O. J. (2020). Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*, *78*(3), 554-565.e7. https://doi.org/10.1016/j.molcel.2020.03.003

Lee, C. (2022). Towards the Genetic Architecture of Complex Gene Expression Traits: Challenges and Prospects for eQTL Mapping in Humans. *Genes*, *13*(2), 235. https://doi.org/10.3390/GENES13020235

Lewis, C. M., & Vassos, E. (2020). Polygenic risk scores: From research tools to clinical instruments. In *Genome Medicine* (Vol. 12, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-020-00742-5

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1–2), 84–98. https://doi.org/10.1016/j.cell.2011.12.014

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., Song, Z., Ji, W., Wang, M., Zhou, J., Chen, B., Liu, Y., Wang, J., Wang, P., Yang, P., … Shi, Y. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nature Genetics*, *49*(11), 1576–1583. https://doi.org/10.1038/NG.3973,

Liu, N., Low, W. Y., Alinejad-Rokny, H., Pederson, S., Sadlon, T., Barry, S., & Breen, J. (2021). Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics & Chromatin 2021 14:1*, *14*(1), 1–17. https://doi.org/10.1186/S13072-021-00417-4

Liu, T., Porter, J., Zhao, C., Zhu, H., Wang, N., Sun, Z., Mo, Y. Y., & Wang, Z. (2019). TADKB: Family classification and a knowledge base of topologically associating domains. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-5551-2

Liu, Z., Liu, R., Gao, H., Jung, S., Gao, X., Sun, R., Liu, X., Kim, Y., Lee, H. S., Kawai, Y., Nagasaki, M., Umeno, J., Tokunaga, K., Kinouchi, Y., Masamune, A., Shi, W., Shen, C., Guo, Z., Yuan, K., … Huang, H. (2023). Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nature Genetics 2023 55:5*, *55*(5), 796–806. https://doi.org/10.1038/s41588-023-01384-0

Long, H. S., Greenaway, S., Powell, G., Mallon, A. M., Lindgren, C. M., & Simon, M. M. (2022). Making sense of the linear genome, gene function and TADs. *Epigenetics and Chromatin*, *15*(1), 1–19. https://doi.org/10.1186/S13072-022-00436-9/FIGURES/6

Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, *11*(1), 1–3. https://doi.org/10.1038/S41467-020-19653-5;TECHMETA=141,43,45;SUBJMETA=205,208,2138,631,692,699;KWRD=DISEASES,GENOME-WIDE+ASSOCIATION+STUDIES

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. In *Nature* (Vol. 461, Issue 7265, pp. 747–753). https://doi.org/10.1038/nature08494

McArthur, E., & Capra, J. A. (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of Human Genetics*, *108*(2), 269–283. https://doi.org/10.1016/j.ajhg.2021.01.001

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, *28*(5), 495–501. https://doi.org/10.1038/NBT.1630,

Mendelian inheritance - Wikipedia. (n.d.). Retrieved May 25, 2025, from https://en.wikipedia.org/wiki/Mendelian_inheritance

Mocellin, S. (2007). Microarray Technology and Cancer Gene Profiling. *Advances in Experimental Medicine and Biology*, *593*. https://doi.org/10.1007/978-0-387-39978-2/COVER

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47–60. https://doi.org/10.1109/79.543975

Morrison, O., & Thakur, J. (2021). Molecular complexes at euchromatin, heterochromatin and centromeric chromatin. *International Journal of Molecular Sciences*, *22*(13). https://doi.org/10.3390/ijms22136922

Mudge, J. M., Carbonell-Sala, S., Diekhans, M., Martinez, J. G., Hunt, T., Jungreis, I., Loveland, J. E., Arnan, C., Barnes, I., Bennett, R., Berry, A., Bignell, A., Cerdán-Vélez, D., Cochran, K., Cortés, L. T., Davidson, C., Donaldson, S., Dursun, C., Fatima, R., … Frankish, A. (2025). GENCODE 2025:

reference gene annotation for human and mouse. *Nucleic Acids Research*, *53*(D1), D966–D975. https://doi.org/10.1093/NAR/GKAE1078,

Nakahara, S., Male, A. G., Turner, J. A., Calhoun, V. D., Lim, K. O., Mueller, B. A., Bustillo, J. R., O'Leary, D. S., Voyvodic, J., Belger, A., Preda, A., Mathalon, D. H., Ford, J. M., Guffanti, G., Macciardi, F., Potkin, S. G., & Van Erp, T. G. M. (2023). Auditory oddball hypoactivation in schizophrenia. *Psychiatry Research - Neuroimaging*, *335*. https://doi.org/10.1016/j.pscychresns.2023.111710

Naqvi, S., Hoskens, H., Wilke, F., Weinberg, S. M., Shaffer, J. R., Walsh, S., Shriver, M. D., Wysocka, J., & Claes, P. (2022). Decoding the Human Face: Progress and Challenges in Understanding the Genetics of Craniofacial Morphology. *Annual Review of Genomics and Human Genetics*, *23*, 383. https://doi.org/10.1146/ANNUREV-GENOM-120121-102607

Ng, S. K. (2013). Recent developments in expectation-maximization methods for analyzing complex data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 415–431. https://doi.org/10.1002/WICS.1277;WEBSITE:WEBSITE:WIRES;WGROUP:STRING:PUBLICATION

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., CHopewell, J., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., MArmasu, S., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., … Farrall, M. (2015). A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics 2015 47:10*, *47*(10), 1121–1130. https://doi.org/10.1038/ng.3396

Nygaard, P., & Saxild, H. H. (2009). Nucleotide Metabolism. *Encyclopedia of Microbiology, Third Edition*, 296–307. https://doi.org/10.1016/B978-012373944-5.00082-1

Ong, T., & Ramsey, B. W. (2023). Cystic Fibrosis: A Review. *JAMA*, *329*(21), 1859–1871. https://doi.org/10.1001/JAMA.2023.8120,

(PDF) Doing database design with MySQL. (n.d.). Retrieved May 26, 2025, from https://www.researchgate.net/publication/271910489_Doing_database_design_with_MySQL

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y., & Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature 2010 464:7289*, *464*(7289), 768–772. https://doi.org/10.1038/nature08872

Pullin, J. M., & Wallace, C. (2024). Variant-specific priors in colocalisation analysis. *BioRxiv*, 2024.08.21.608957. https://doi.org/10.1101/2024.08.21.608957

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/BIOINFORMATICS/BTQ033,

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/NAR/GKZ369

Rocha, P. P., Raviram, R., Bonneau, R., & Skok, J. A. (2015). Breaking TADs: Insights into hierarchical genome organization. *Epigenomics*, *7*(4), 523–526. https://doi.org/10.2217/EPI.15.25

Schoenfelder, S., & Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics 2019 20:8*, *20*(8), 437–455. https://doi.org/10.1038/s41576-019-0128-0

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). DbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/NAR/29.1.308,

Shu, X., Long, J., Cai, Q., Kweon, S. S., Choi, J. Y., Kubo, M., Park, S. K., Bolla, M. K., Dennis, J., Wang, Q., Yang, Y., Shi, J., Guo, X., Li, B., Tao, R., Aronson, K. J., Chan, K. Y. K., Chan, T. L., Gao, Y. T., … Zheng, W. (2020). Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/S41467-020-15046-W;TECHMETA=43,45;SUBJMETA=1347,631,67,69;KWRD=BREAST+CANCER,CANCER+GENOMICS

Spitz, F. (2016). Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. In *Seminars in Cell and Developmental Biology* (Vol. 57, pp. 57–67). Academic Press. https://doi.org/10.1016/j.semcdb.2016.06.017

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., & Lancet, D. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*, *2016*, 1.30.1-1.30.33. https://doi.org/10.1002/CPBI.5,

Strome, S., Bhalla, N., Kamakaka, R., Sharma, U., & Sullivan, W. (2024). Clarifying Mendelian vs non-Mendelian inheritance. In *Genetics* (Vol. 227, Issue 3). Oxford University Press. https://doi.org/10.1093/genetics/iyae078

Sundd, P., Gladwin, M. T., & Novelli, E. M. (2019). Pathophysiology of Sickle Cell Disease. *Annual Review of Pathology: Mechanisms of Disease*, *14*, 263–292. https://doi.org/10.1146/ANNUREV-PATHMECHDIS-012418-012838,

Szabo, Q., Bantignies, F., & Cavalli, G. (2019). *Principles of genome folding into topologically associating domains*. https://www.science.org

Tomczak, A., Mortensen, J. M., Winnenburg, R., Liu, C., Alessi, D. T., Swamy, V., Vallania, F., Lofgren, S., Haynes, W., Shah, N. H., Musen, M. A., & Khatri, P. (2018). Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Scientific Reports 2018 8:1*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-23395-2

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. In *Nature Reviews Methods Primers* (Vol. 1, Issue 1). Springer Nature. https://doi.org/10.1038/s43586-021-00056-9

Umans, B. D., Battle, A., & Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends in Genetics*, *37*(2), 109–124. https://doi.org/10.1016/j.tig.2020.08.009

Valdar, W., Solberg, L. C., Gauguier, D., Cookson, W. O., Rawlins, J. N. P., Mott, R., & Flint, J. (2006). Genetic and environmental effects on complex traits in mice. *Genetics*, *174*(2), 959–984. https://doi.org/10.1534/genetics.106.060004

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Võsa, U., Claringbould, A., Westra, H. J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M. J., Agbessi, M., … Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, *53*(9), 1300–1310. https://doi.org/10.1038/S41588-021-00913-Z;SUBJMETA=199,200,205,208,2138,631;KWRD=GENE+EXPRESSION,GENE+REGULATION,GENOME-WIDE+ASSOCIATION+STUDIES

Watanabe, K., Taskesen, E., Van Bochoven, A., & Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/S41467-017-01261-5;SUBJMETA=114,205,208,2138,2401,631;KWRD=DATA+INTEGRATION,GENOME-WIDE+ASSOCIATION+STUDIES

Wei, W. H., Hemani, G., & Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics 2014 15:11*, *15*(11), 722–733. https://doi.org/10.1038/nrg3747

What's New In Python 3.11 — Python 3.13.3 documentation. (n.d.). Retrieved May 26, 2025, from https://docs.python.org/3/whatsnew/3.11.html

Wu, T. (1993). An Accurate Computation of the Hypergeometric Distribution Function. *ACM Transactions on Mathematical Software (TOMS)*, *19*(1), 33–43. https://doi.org/10.1145/151271.151274;TAXONOMY:TAXONOMY:ACM-PUBTYPE;PAGEGROUP:STRING:PUBLICATION

Yang, J. H., & Hansen, A. S. (2024). Enhancer selectivity in space and time: from enhancer–promoter interactions to promoter activation. In *Nature Reviews Molecular Cell Biology* (Vol. 25, Issue 7, pp. 574–591). Nature Research. https://doi.org/10.1038/s41580-024-00710-6

Yap, C. X., Lloyd-Jones, L., Holloway, A., Smartt, P., Wray, N. R., Gratten, J., & Powell, J. E. (2018). Trans-eQTLs identified in whole blood have limited influence on complex disease biology. *European Journal of Human Genetics 2018 26:9*, *26*(9), 1361–1368. https://doi.org/10.1038/s41431-018-0174-7

Zhou, X., & Cai, X. (2021). Joint eQTL mapping and inference of gene regulatory network improves power of detecting both cis- and trans-eQTLs. *Bioinformatics*, *38*(1), 149. https://doi.org/10.1093/BIOINFORMATICS/BTAB609

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/NG.3538;TECHMETA=41,43,45,47;SUBJMETA=199,205,208,2138,631;KWRD=GENE+EXPRESSION,GENOME-WIDE+ASSOCIATION+STUDIES

Zou, H., Wu, L. X., Tan, L., Shang, F. F., & Zhou, H. H. (2020). Significance of Single-Nucleotide Variants in Long Intergenic Non-protein Coding RNAs. In *Frontiers in Cell and Developmental Biology* (Vol. 8). Frontiers Media S.A. https://doi.org/10.3389/fcell.2020.00347

Use of Generative Artificial Intelligence (GenAI) – Form to be completed

Student name: D'hoore Alita
Student number: r1009030

Please indicate with "X" whether it relates to a course assignment, to the BIG-project or to the master's thesis:

⊗ This form is related to my master's thesis.
Title master's thesis: Novel method for annotating gene targets to GWAS loci
Promoter: Prof. Peter Claes

O This form is related to a BIG-project.
Title BIG-project: …
Promoter: …

O This form is related to a course assignment.
Course name: …
Course code: …

Please indicate with "X":

- o I did not use GenAI tools.
- ⊗ I did use GenAI tools. In this case specify which one (e.g. ChatGPT/GPT4/...): ChatGPT

Please indicate with "X" (possibly multiple times) in which way you were using it:

- ⊗ **As a language assistant for reviewing or improving texts you wrote yourself, provided that the model does not add new content.** In this case, the use of GenAI is similar to the spelling and grammar check tools we already have today, so you do not need to explicitly mention using GenAI for this).
- ⊗ **As a search engine to get initial information on a topic or to make an initial search for existing research on the topic.** (This way of gathering information is similar to using an ordinary search engine when working on an assignment. As a student, you are responsible for checking and verifying the absence and correctness of references. Therefore, after this initial search, look for scientific sources and conduct your own analysis of the source documents. Interpret, analyze and process the information you obtained; don't just copy-paste it. If you then write your own text based on this information, you do not have to mention you used GenAI.)
- o **To generate text blocks.** (If you do copy-paste text blocks of GenAI output, you have to cite your GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)
- o **To generate graphs or figures.** (If you do copy-paste graphs/figures of GenAI output, you have to cite the GenAI sources and quote them, i.e. you clearly state that the item was created via GenAI by citation/reference.)
- ⊗ **To generate some code as part of a larger assignment.** (Watch out, this can only be done if the teacher/promotor explicitly allows it.)
- o **Other** (Contact the teacher of the course or the supervisor of the thesis or BIG project. Explain how you comply with article 84 of the examination regulations. Explain the usefulness or added value of using GenAI.)
- o

Further important guidelines and remarks:

The faculty follows the KU Leuven policy regarding responsible use of GenAI. This form is an aid towards transparency about the use of GenAI by the student which is essential. Irresponsible and non-transparent use of GenAI can be considered an irregularity and can be sanctioned. Students who consider to use GenAI should inform themselves through the university website concerning the additional guidelines (How to correctly quote and refer to GenAI? What is (not) allowed? Tips and points of attention for responsible use):
https://www.kuleuven.be/english/education/student/educational-tools/generative-artificial-intelligence