RetroFun-RVS: a retrospective family-based framework for rare variant analysis incorporating functional annotations

Loic Mangnier[1,2,3], Alexandre Bureau[1,2,3*]

[1] Social and Preventive Medicine Department, Laval University, Quebec City, QC, Canada [2] CERVO Brain Research Center, Quebec City, QC, Canada [3] Big Data Research Center, Quebec City, QC, Canada

[*] corresponding author: alexandre.bureau@fmed.ulaval.ca

Running Title: USING FUNCTIONAL ANNOTATIONS IN FAMILY-BASED RARE VARIANT ANALYSIS

Keywords: Non-coding genome, Pedigree-based association tests, Variant sharing, 3D genome

1 Abstract

A large proportion of genetic variations involved in complex diseases are rare and located within non-coding regions, making the interpretation of underlying biological mechanisms difficult. Although technical and methodological progresses have been made to annotate the genome, current disease- rare-variant association tests incorporating such annotations suffer from two major limitations. Firstly, they are restricted to case-control designs of unrelated individuals, which often require tens or hundreds of thousands of individuals to achieve sufficient power. Secondly, they were not evaluated with region-based annotations needed to interpret the causal regulatory mechanisms. In this work we propose RetroFun-RVS, a new retrospective family-based score test, incorporating functional annotations. One of the critical features of the proposed method is to aggregate genotypes while measuring rare variant sharing among affected family members to compute the test statistic. Through extensive simulations, we have demonstrated that RetroFun-RVS integrating networks based on 3D genome contacts as functional annotations reaches greater power over the region-wide test, others strategies to include sub-regions and competing methods. Also, the proposed framework shows robustness to non-informative annotations, keeping a stable power when causal variants are spread across regions. We provide recommendations when dealing with different types of annotations or family structures commonly encountered in practice. In summary we argue that RetroFun-RVS, by allowing integration of functional annotations corresponding to regions or networks with transcriptional impacts, is a useful framework to highlight regulatory mechanisms involved in complex diseases.

2 Introduction

Over the past few years with the democratization of whole-exome or whole-genome sequencing data, important progresses have been made in the effort to link genetic variations to phenotypes. Indeed, at population scale, Genome-Wide Association Studies (GWAS) have provided useful resources to highlight variants involved in diseases. However, these methods, in addition to requiring tens or hundreds of thousands of individuals, are mainly restricted to common variants, leaving an important part of heritability unexplained (Manolio et al., 2009). In fact, studies have shown that the individual genetic risk is also influenced by rare variants (minor allele frequency (MAF) $\leq 1\%$), (Singh et al., 2022; Sun et al., 2022). In addition to being rare, variants influencing disease risk tend to be located within non-coding regions, making the underlying biological mechanisms difficult to interpret (Zhang and Lupski, 2015). Thus, the tremendous amount of rare variants located within non-coding regions brings new challenges to identify new causal variants involved in diseases, and accounting for their functional impacts remains crucial from a fine-mapping perspective (Schaid et al., 2018).

Methods have been proposed to overcome the challenge of sparsity. Indeed, because variants are rare, methods testing them in an unitary fashion perform badly (Madsen and Browning, 2009). Thus, rare-variants association tests (RVATs) are methods aggregating genotypes across several variant sites within a gene, pathway or regions functionally close. By collapsing variants across over regions, these methods considerably reduce the number of tests throughout the genome, hence increasing statistical power. Among them, burden tests were initially proposed and are powerful when all variants across regions show a homogeneous effect (B. Li and Leal, 2008; Madsen and Browning, 2009). However, when regions combine both deleterious and protective variants, burden tests comparing cases to controls suffer from a substantial decrease of power. Alternatives to address this limitation have been proposed (Ionita-Laza et al., 2011; Neale et al., 2011). One of the critical features of RVATs is that they can be expressed through regression models, allowing the integration of variant weights, either fixed (based on the MAF), or estimated in a data adaptive manner (Madsen and Browning, 2009; M. C. Wu et al., 2011). The multiple ways to define test statistics created a need for combining several p-values within a given region to assess the association with a trait, while adjusting for multiplicity. Liu et al., 2019 have proposed ACAT, a powerful statistical framework combining p-values in an efficient way. One of the major advantages of ACAT over other combination methods (e.g., taking the minimum (minP), Fisher's method) is that it requires neither resampling procedures, nor independent p-values nor explicit models for correlations. Although these set-based tests have made possible the discovery of new regions involved in complex diseases, they required very large sample sizes.

An alternative approach to samples of unrelated subjects is to exploit family-based studies. In addition to reducing genetic heterogeneity, pedigree-based studies have been shown to have more power than population-based approaches

for detecting rare variants, when an enrichment of risk variants among families is expected (Laird and Lange, 2006; M. Li et al., 2006; Ott et al., 2011). Information provided by variants segregating with the disease, even imperfect, can be exploited to highlight new causal variants, given a second breath to studies in extended pedigrees (e.g., Bureau et al., 2014). Recent methods based on identity-by-descent (IBD) or combining both linkage approaches and RVATs have been developed (Bureau et al., 2019; Sul et al., 2016; Zhao et al., 2019). One important feature of these approaches is that they focus on, or can be restricted to, only affected family members, when these are expected to contribute more information than unaffected subjects (Schaid et al., 2010). Affected-only designs have a long tradition in gene-gene or gene-environment interaction analysis and have been extended to family-based studies, requiring smaller sample sizes to reach equivalent power, compared to considering unrelated case-only individuals, which is an appealing feature in practice (W. Li et al., 2019). However, selecting individuals retrospectively (based on their phenotype) may lead to highly ascertained sampling schemes resulting in overestimated association measures. Retrospective likelihood, by conditioning on phenotype, have been shown to give accurate estimates for common variants when ascertainment bias is expected (Schaid et al., 2010). Extensions of RVATs have been proposed for retrospective samples of families (Schaid et al., 2013). A limitation of all the above methods is that none of them currently integrates external information on biological mechanisms involved in diseases. How to leverage information on non-coding regulatory elements in the detection of variants influencing disease risk remains an open question. Thus, there was an increasing interest in using external information for this task, and hence highlighting the biological mechanisms. Recent methods, such as FST (He et al., 2017) or FunSPU (Y. Ma and Wei, 2019) have proposed to adaptively test functional annotations under a general RVAT framework. These methods have shown substantial increases in power when at least one functional score is predictive for the effect of variants on the trait, while they show robustness when no annotations were predictive for variant impact on the trait, revealing new causal variants involved in complex traits. More recently, with the striking development of methods detecting regulatory elements such as enhancers (Fulco et al., 2019; Yao et al., 2022), progresses have been made in associating non-coding SNPs to their target genes (Gazal et al., 2022; Nasser et al., 2021). Subsequently, some authors have proposed to incorporate this information within statistical frameworks. Hence, S. Ma et al., 2021 have demonstrated that long range 3D interactions between genes and enhancers add information for the integration of non-coding regulatory regions within gene-based frameworks. This model only considers pairs of gene-enhancer, consistent with previous studies (C. Wu and Pan, 2018). Models extending gene-enhancer pairs to Cis-Regulatory Hubs (CRHs), networks encompassing up to several genes and active enhancers have been proposed (Mangnier et al., 2022). CRHs have been shown to be a relevant model in schizophrenia etiology, explaining more heritability than tissue- and non-tissue- specific elements, and being more effective to link noncoding SNPs to differentially expressed genes in schizophrenia compared to Topologically As-

sociated Domains (TADs) or pairs of gene-enhancer. To our knowledge, no study to date has proposed to integrate functional annotations within a RVAT framework, while allowing the incorporation of discontinuous genomic regions involved in 3D-based networks and exploiting family-based designs.

In this paper, we propose RetroFun-RVS (Retrospective Functional Rare Variant Sharing), a model, allowing the integration of functional annotations under a family-based design considering only affected individuals. Through extensive simulation studies, we have demonstrated that RetroFun-RVS integrating CRHs as functional annotations is a more powerful approach to detect causal variants over others strategies, while well controlling the Type I error rate. We provided recommendations when dealing with different types of functional scores or pedigree structures. Finally, we have demonstrated that integrating 3D-based functional annotations through networks is a relevant strategy to gain power of detection of causal variants, while highlighting the underlying biological mechanisms involved in diseases.

3 Material and Methods

3.1 Notations and Model

Suppose that we have N subjects within F families, where n_f is the number of individuals for the fth family. Let's define, Y, a binary vector of phenotypes, G a $N \times p$ matrix of genotypes for rare variants, coded as the number of copies of the minor allele 0, 1, 2. Assuming a log-additive model for the individual SNP effect on disease risk, under assumption of conditional independence of the phenotypes of different individuals given their genotypes and considering only affected individuals, following Schaid et al., 2010, the retrospective likelihood for one family can be written as:

$$P(G|Y) = \frac{\exp \sum_{i \in D} \sum_{j=1}^{p} \beta_j x_{ij} P(G)}{\sum_{G^*} \exp \sum_{i \in D} \sum_{j=1}^{p} \beta_j x^*_{ij} P(G^*)}$$

where D is the subset of affected members in the family, while x_{ij} is a condensed notation for $x(G_{ij})$, the number of minor alleles for variant j in individual i in the multilocus genotype configuration G. We make the assumption that only one copy of the minor allele was introduced once by a family founder, implying x_{ij} can only take the values 0 or 1 in the absence of inbreeding in the family. The software implementation of RetroFun-RVS converts genotypes homozygous for the rare allele to heterozygous genotypes by default (i.e., $x_{ij} = 2$ is changed to $x_{ij} = 1$). An alternative option is to discard variants with homozygous rare genotypes. In Schaid et al., 2010, $P(G)$ is the unconditional genotype probability and depends on MAF, which needs to be estimated in practice. Instead, we opted for conditioning the probability on the event of observing at least one a copy of each RV j present in the family (i.e., $\sum_i x_{ij} \geq 1$) as in Bureau et al., 2019. In addition, we combined this conditional probability with the assumption that the variant frequency tends to 0, hence the probability does

not depend on MAF and therefore the computation does not require external variant frequency estimates. In this context, the genotypes can be interpreted as rare variant sharing patterns, hence RVS in the method name. The sum in the denominator is over all genotype configurations respecting the condition within the given pedigree, where G^* denotes one particular configuration. Since we expect that risk variant effects dominate protective variants in the score test statistic when considering only affected individuals (Supplementary methods and Figure S1), we propose to adapt the retrospective framework for a burden test (B. Li and Leal, 2008; Madsen and Browning, 2009). So, we can express β_j the effect of the jth variant through $\beta_0 w_j$ where w_j is usually a weighting function to specify variant effects through a function of MAF. From now on and in the following sections we will consider $wj = \beta(MAF_j, 1, 25)$ to up-weights rare variants.

As suggested by He et al., 2017, the effect for the jth variant can be partitioned with respect to functional annotations $Z_{jk}, k = 1 \ldots q$. So, for the Burden this leads to:

$$\beta_0 = \sum_j w_j \sum_{k=0}^{q} Z_{jk}\gamma_k$$

with $Z_{j0} = 1$ and γ_0 corresponding to the original burden test parameter. Intuitively, partitioning variant effect allows a modulation of the variant effect based on MAF and functional annotations. Moreover, γ_0 ensures a minimal power when no predictive functional annotations are present for the trait. When at least one annotation is predictive, the partitioned model offers increased power over the original test (He et al., 2017).

Now combining the retrospective likelihood model described by Schaid et al., 2010 and the decomposed variant effect, we obtain:

$$P(G|Y) = \frac{\exp(\sum_{i\in D}\sum_{j=1}^{p} w_j x_{ij} \sum_{k=0}^{q} Z_{jk}\gamma_k)P(G)}{\sum_{G^*}\exp(\sum_{i\in D}\sum_{j=1}^{p} w_j x^*_{ij} \sum_{k=0}^{q} Z_{jk}\gamma_k)P(G^*)}$$

Thus for the kth functional annotation the score function summed across the F families is :

$$S_k(\gamma_k) = \sum_{f=1}^{F}\left(\sum_{j=1}^{p} w_j Z_{jk}\left(\sum_{i\in D} x_{fij} - \frac{\sum_{G^*_f}\sum_{i\in D} x^*_{fij}\exp(\sum_{j=1}^{p} w_j Z_{jk}\gamma_k \sum_{i\in D} x^*_{fij})P(G^*_f)}{\sum_{G^*_f}\exp(\sum_{j=1}^{p} w_j Z_{jk}\gamma_k \sum_{i\in D} x^*_{fij})P(G^*_f)}\right)\right)$$

Intuitively, this quantity can be seen as the difference between the observed genotype value and the expected value, weighted by MAF and functional annotations. Setting γ_k to 0, we obtain the score statistic:

$$S_k(0) = \sum_{f=1}^{F}\left(\sum_{j=1}^{p} w_j Z_{jk}\left(\sum_{i\in D} x_{fij} - \sum_{G^*_{fj}}\sum_{i\in D} x^*_{ij}P(G^*_{fj})\right)\right)$$

The genotype probability required $P(G_{fj})$ is for a single variant configuration in family f and can be computed using RVS (Sherman et al., 2019). $S_k^2(0)$ asymptotically follows a ${\chi_1}^2$, when properly scaled by the variance of $S_k(0)$. This variance can be obtained by combining sharing patterns and observed genotypes within families. Moreover, simplifications may be obtained from assumptions on the linkage disequilibrium structure (See Supplementary methods). However, we observed when only few variants are expected within a functional annotation that resampling procedures may be required to control the Type I error rate. We proposed to resample observed variant counts based on family-specific genotype configuration probabilities $P(G_{fj})$. Thus, this bootstrap procedure maintains both the linkage disequilibrium and family structures of the data.

For testing multiple functional scores within a single unified test $H_0 : \forall k, \gamma_k = 0$ vs $H_1 : \exists k, \gamma_k > 0$, we then propose to combine $q + 1$ single p-values corresponding to the q functional annotations and the original burden with ACAT (Liu et al., 2019). Briefly, ACAT aggregates individual p-values and approximates the test statistic (and the subsequent p-value) based on a Cauchy distribution. So, for $q + 1$ tests in a region of interest, the ACAT statistic can be written as:

$$T_{ACAT} = \sum_{k=0}^{q} tan((0.5 - p_k)\pi)$$

4 Numerical Simulations

Genotypes were simulated based on observed variant sites and their corresponding MAF for the European population from the 1000 Genome Project database (phase 3). We extracted the 510 rare (MAF $\leq 1\%$) coding non-synonymous and within-enhancer non-coding single nucleotide variants from a region of 800Kb (chr1:24100000-24970000), corresponding to a TAD in iPSC-derived neurons. This TAD has been selected since it encompasses four CRHs showing different complexities (two genes-five enhancers (CRH1); two genes-two enhancers (CRH2); one gene-one enhancer (CRH3); one gene-four enhancers (CRH4); See Table S1). Refer to Mangnier et al., 2022 for more details. Using RarePedsim (B. Li et al., 2015), we generated sequence data for 270 affected subjects in the primary sample of 52 extended and small pedigrees (Figures 1 and S2) and a secondary sample of 81 small pedigrees (Figure S2). For both Type I error rate and power evaluation, we simulated dichotomous phenotypes from a logistic model without covariates, specifying a population prevalence of 1%. Details on pedigree structures and simulations were provided in the Supplementary methods. We focused on evaluating the ACAT-combined p-values.

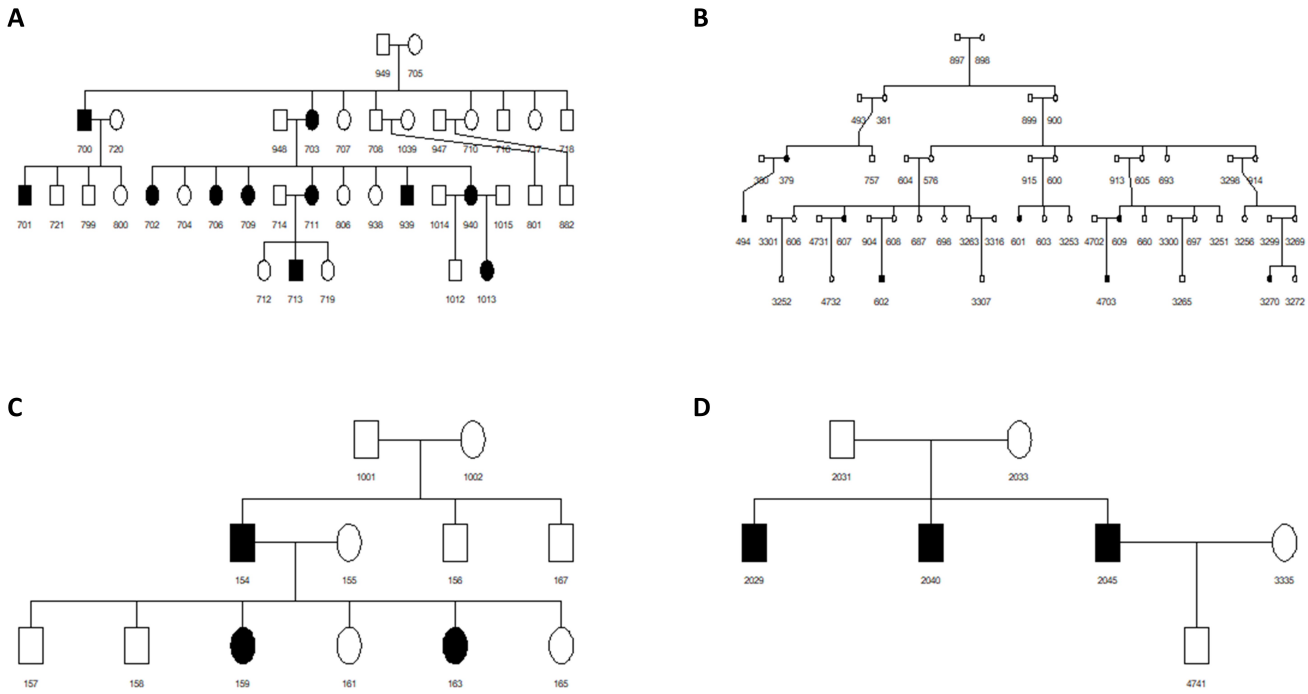

Figure 1: Example of pedigree structures considered in the simulation studies. Affected subjects are indicated by filled squares or circles

4.1 Type I Error Simulations

To determine whether the proposed framework preserves the desired Type I error rate, genotype data were generated unconditional on the affection status for family members. We specified a null effect for variants observed in families, i.e., odds-ratio (OR) = 1. Generating one thousand replicates, we first examined the performance of RetroFun-RVS$_{CRHs}$, which is RetroFun-RVS applied to CRHs and including variants over the entire TAD as global burden, with alternative definitions of regions to be included as functional scores: RetroFun-RVS$_{Pairs}$, RetroFun-RVS$_{Genes}$, and RetroFun-RVS$_{Sliding-Window}$, for the method considering pairs of gene-enhancers, genes and a 10 Kb sliding window, respectively (Figure 2). We also assessed whether the method is well-calibrated in presence of small families. Results for this setting were reported in Supplementary figures.

4.2 Empirical Power Simulations

We set 2% of the variants over the entire region to be risk variants as suggested before S. Ma et al., 2021, also performing simulations with 1% of risk variants as a sensitivity analysis. Genotypes were generated conditional on the affection status for each pedigree member assuming a multiplicative model with fixed variant effect, i.e., not depending on the MAF. We considered different scenarios where we varied the proportion of causal variants found in CRHs: 100%, 75% and 50% of causal variants (OR=5) were located within one CRH. The remaining variants being neutral (OR=1). This scenario is expected when variants are concentrated within elements functionally close. These three proportions correspond to the most advantageous scenario where all causal variants are within the same region and two mixed scenarios where signal is spread across the sequence of the region at different degrees. Our first evaluation assessed the gain of power by incorporating CRHs as functional annotations over the test including no scores (referred to as Burden Original). We also compared RetroFun-RVS$_{CRHs}$ with others strategies to incorporate regions as functional annotations: RetroFun-RVS$_{Pairs}$, RetroFun-RVS$_{Genes}$, and RetroFun-RVS$_{Sliding-Window}$, for the method considering pairs of gene-enhancers, genes and a 10 Kb sliding window, respectively (Figure 2). Also, we assessed the performance in terms of power of our method compared to existing approaches namely, RVS (Bureau et al., 2019) and RV-NPL (Zhao et al., 2019) (Figure S3). Power was evaluated as the proportion of p-values less than α = 8.33e-6, corresponding to the Bonferroni-adjusted 0.05 significance level when testing six thousand independent regions across the genome, corresponding to three thousand TADs (the average number of TADs found in our previous study across cell-types or tissues (Mangnier et al., 2022)), while permitting the same number of additional domains of interest, i.e., outside TADs, to be tested. Results at lower proportion of risk variants and considering small pedigrees were given in the Supplements.

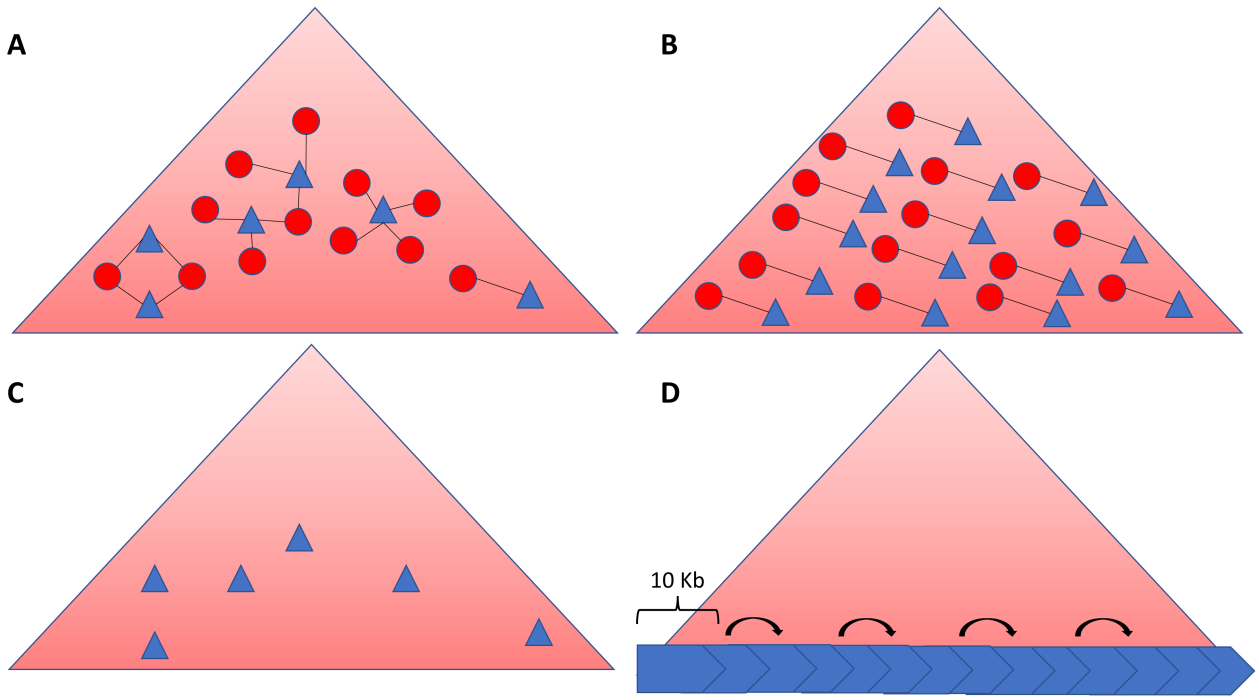

Figure 2: Overview of functional annotations considered in the simulation studies. For all the 4 panels, big red triangles represent the selected TAD for the simulation studies, small blue triangles the genes (exons + promoters), and red circles the enhancers. (**A**) CRHs as functional annotations. (**B**) Pairs as functional annotations. CRHs are split with respect to each gene-enhancer pair. (**C**) Genes as functional annotations. (**D**) 10 Kb sliding windows as functional annotations.

5 Results

5.1 Simulation of Type I Error Rate

The results show that, when we considered CRHs as functional annotations, the Type I error rate was slightly conservative for modest p-values while well-controlled for more stringent thresholds, except for the extreme smallest p-value (Figure 3). However, when we considered variants as independent in the variance calculation, Type I error rate was slightly inflated (Figure S4). Moreover, RetroFun-RVS with no functional annotation (i.e., a single test of all RVs in the TAD) was conservative when we considered variant dependence through covariance terms in the variance calculation, while it was well-calibrated assuming variant independence (Figure S5). Results for RetroFun-RVS$_{CRHs}$ for each individual score show that the approach with covariance terms is either well cal-

ibrated or slightly conservative (Figure S6-S8). In addition, the method shows moderate Type I error rate inflation when applied to small family structures, increasing when assuming variant independence (Figure S9). Furthermore, integrating pairs and genes as functional annotations, we observed moderate inflation of the Type I error rate in extended pedigrees, even when considering variant dependence, while for 10Kb sliding windows the Type I error rate inflation was severe (Figure S10). We attempted to discard 10 kb windows with few variants, and observed that Type I error control was achieved on windows encompassing 30 variants or more (results not shown). Moreover, the bootstrap procedure applied to RetroFun-RVS$_{Pairs}$, RetroFun-RVS$_{Genes}$ and , RetroFun-RVS$_{Sliding-Window}$ to compute p-values empirically provides good Type I error rate control, while slightly conservative, even for functional annotations encompassing few variants (Figure S11). To summarize, the results show that RetroFun-RVS with asymptotic p-values is a valid approach when CRHs or a large region are considered in extended pedigrees, despite being inflated to various degrees for others strategies or family structures. Bootstrap p-values can be computed in these instances to control the Type I error rate.

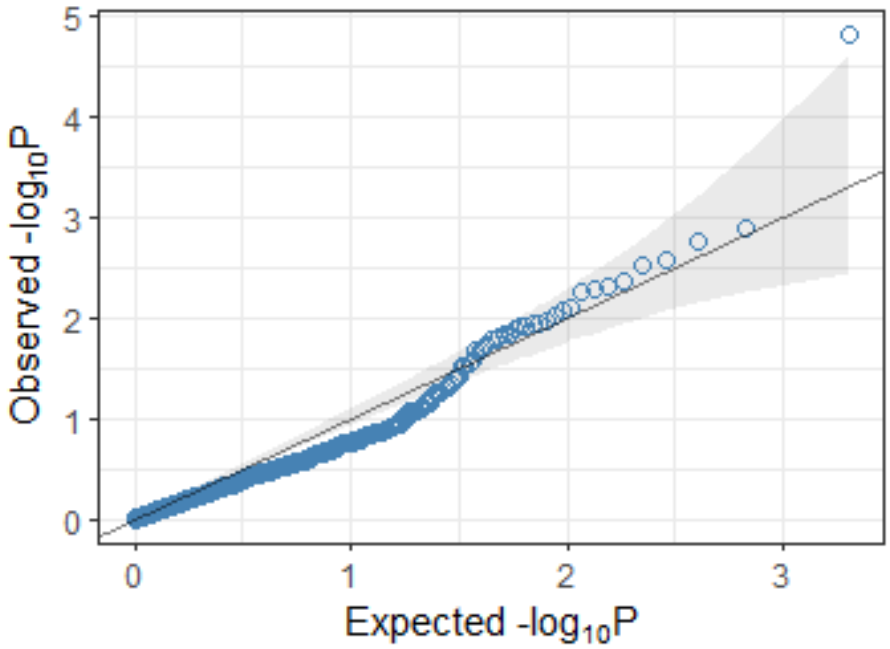

Figure 3: Quantile-Quantile plot of ACAT-Combined P-values for RetroFun-RVS$_{CRHs}$ considering variant dependence. Because only a few replicates had p-values for CRH 3, we omitted it in the analysis.

5.2 Power Comparison Considering Different Strategies to build Functional Annotations

In the first set of power evaluations, we assessed power under different scenarios of causal variant distributions. Firstly, we compared RetroFun-RVS integrating CRHs with the same method incorporating no functional annotation. Consequently, when 100% and 75% of causal variants were within one CRH, our method RetroFun-RVS$_{CRHs}$ performed better than the original burden test showing gains of 10% and 11%, while at 50% causal the power remains comparable (Figure 4A). Also, considering only small pedigrees, we observed that, even if both RetroFun-RVS$_{CRHs}$ and the original burden test without annotation exhibit lower power, the gain for RetroFun-RVS$_{CRHs}$ becomes higher as the percentage of causal variant within the CRH of interest increases (Figure S12). Congruent results were obtained when a lower proportion of causal variants was considered, showing a minimal power gain of 16% (Figure S13). Therefore, our findings suggest that substantial power gain can be achieved when CRHs are predictive for the effect of variants on the trait, RetroFun-RVS$_{CRHs}$ showing robustness when signal is spread across several CRHs. Then, we compared RetroFun-RVS$_{CRHs}$ to other strategies to integrate regions as functional annotations, namely RetroFun-RVS$_{pairs}$, RetroFun-RVS$_{genes}$, and RetroFun-RVS$_{Sliding-Window}$. Our results show that integrating CRHs as functional annotations is a more powerful strategy compared to the other strategies considered 4B). Globally our results follow the same pattern when decreasing the proportion of causal variants (Figure S14). In summary, RetroFun-RVS$_{CRHs}$ exhibits gain of powers when CRHs show high or modest percentages of causal variants. Also, the method is robust and powerful under the different scenarios that we considered, that are, inclusion of weakly predictive CRHs, small percentages of risk variants, and presence of small families.

5.3 Power Comparison with Others Affected-Only Methods

In the second set of power evaluations, we compared RetroFun-RVS$_{CRHs}$ with others affected-only methods, namely RVS (Bureau et al., 2019) and RV-NPL (Zhao et al., 2019). Thus, to proceed to fair comparisons between methods, we adapted RVS and RV-NPL to take CRHs into account (See Supplementary methods). With 2% risk variants, when we considered 75% of causal variants located within one CRHs, we observed that RetroFun-RVS reaches greater power compared to competing methods 4C), exhibiting significantly shorter computation times (Table 1). At lower proportions of risk variants, the new method remains more powerful compared to RV-CHP or RVS, and equivalent to RV-NPL (Figure S15).

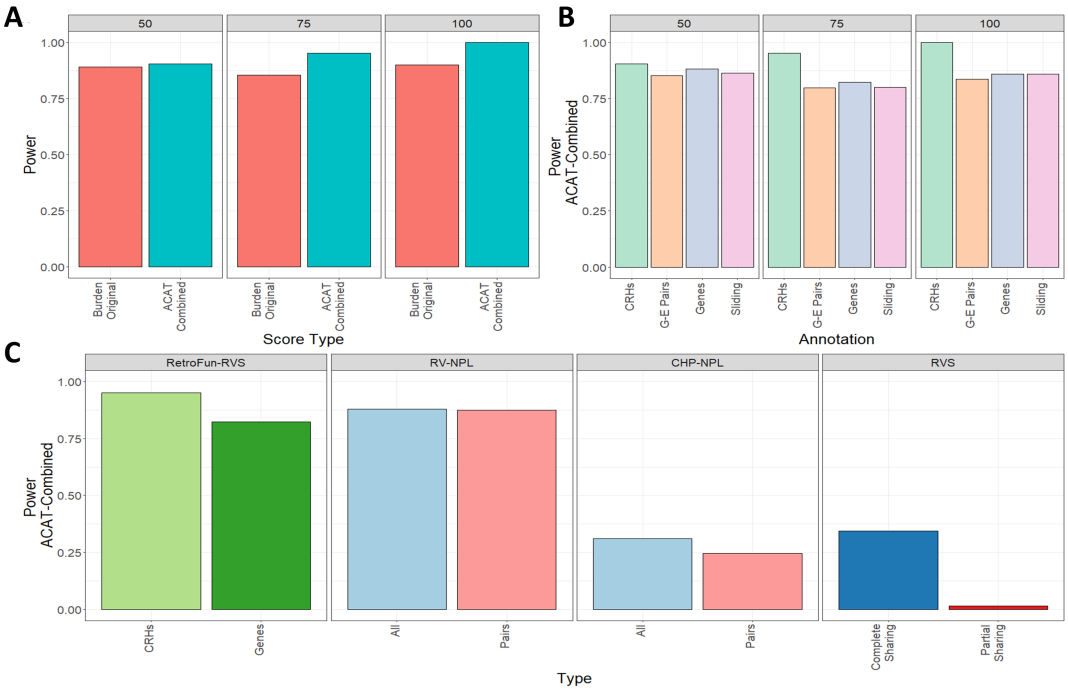

Figure 4: Power evaluation of RetroFun-RVS under different scenarios for 2% risk variants. (**A**) Power at different proportions of risk variants within the CRH, between RetroFun-RVS$_{CRHs}$ with no functional annotation (Burden Original) and RetroFun-RVS$_{CRHs}$ including the four CRHs (ACAT-Combined). Power was evaluated on the basis of 1,000 replicates. (**B**) Power at different proportions of risk variants within the CRH between RetroFun-RVS$_{CRHs}$ (CRHs), RetroFun-RVS$_{Pairs}$ (G-E Pairs), RetroFun-RVS$_{Genes}$ (Genes), and RetroFun-RVS$_{Sliding-Window}$ (Sliding). To correct Type I error inflation observed in RetroFun-RVS$_{Sliding-Window}$, we only considered windows encompassing 30 variants or more. Power was evaluated on the basis of 1,000 replicates. (**C**) Power at 75% risk variants within one CRH between RetroFun-RVS$_{CRHs}$ and other affected-only competing methods. Here we included RetroFun-RV$_{genes}$ to mimic CHP-NPL procedure. Power for RetroFun-RVS$_{CRHs}$ and RetroFun-RVS$_{Genes}$ was evaluated on the basis on 1,000 replicates, while for RV-NPL and RVS we generated 200 replicates.

RetroFun-RVS				RV-NPL	CHP-NPL	RVS	
CRHs	G-E Pairs	Genes	Sliding	All + Pairs	All + Pairs	Complete	Partial
1.06	2.02	1.11	11.05	971.4	1823.4	14.26	443.5

Table 1: Running times (in seconds) for analyzing rare variants in the TAD, in one simulated replicate, using a single 2.10GHz processor. For RV-NPL empirical p-values were obtained based on 1 million permutations.

6 Discussion

Most of rare genetic variations are located within non-coding regions, making the underlying biological mechanisms through which they impact disease risk difficult to interpret. Over the past few years, efforts were not only made in annotating the genome but also integrating these annotations into statistical frameworks (He et al., 2017; Y. Ma and Wei, 2019). Although such methods have already been developed for unrelated subjects such as case-control samples, to our knowledge, no approach to date has been proposed to integrate functional annotations within family-based designs. In this paper we have presented RetroFun-RVS, a retrospective burden test, integrating functional annotations considering only affected individuals within families. We have shown that binary annotations corresponding to disjoint regions with regulatory impacts, such as CRHs, provide power gains when such regions concentrate causal variants, outperforming other strategies or competing methods (Figure 4), while well controlling the Type I error rate (Figure 3). Since regulatory mechanisms are highly tissue- or context-dependent it can be challenging to have the right tissue for the right trait, and misspecifying the model is likely in practice. Thus, integrating the original burden test in RetroFun-RVS makes it robust, showing stable power when functional annotations poorly predict the trait. Finally, by computing p-values asymptotically, RetroFun-RVS is computationally faster than competing methods, which often require permutation-based approaches or exact probability computations to sharply control the Type I error rate.

The main rationale behind RetroFun-RVS is that risk variants are enriched among affected individuals compared to the expected variant count based on their relationships. Hence, one critical feature of our method is to aggregate genotypes while measuring rare variant sharing among affected family members to compute the test statistic. However, to implement an affected-only analysis, where individuals are selected based on their disease status, we have adopted a retrospective approach, considering genotypes as random, while conditioning on phenotypes (Schaid et al., 2010). Also, since genotype probabilities do not depend on MAF under the assumption that the variant frequency tends to 0, RetroFun-RVS necessitates only familial information to compute these probabilities, in order to derive the score statistic and its variance (See Material and Methods). This aspect is central, since the variance terms need to be computed only once for the entire set of families, which is computationally efficient even in presence of large pedigrees. Our rare variant assumption however implies

that genotypes homozygous for the rare allele are impossible in the absence of inbreeding. Data simulated for Type I error and power assessments did contain the small number of homozygous rare genotypes expected for variants with MAF = 1%. Conversion to heterozygous genotypes did not increase Type I error rate compared to removing the variants with homozygous rare genotypes (results not shown).

Moreover, RetroFun-RVS in its current form is restricted to binary phenotypes and does not allow the integration of individual-level covariates, such as sex, age or genetic principal components. Hence, future work is needed to extend the framework to continuous phenotypes and include covariates. Also, future works are needed to extend RetroFun-RVS when more than one copy of the minor allele is introduced, which can arise in presence of inbreeding.

In addition to being computationally effective, RetroFun-RVS is more powerful than other affected-only competing methods, under certain scenarios (Figure 4C, Figure S15). For example, compared to RVS, on which RetroFun-RVS is built upon, but which can only analyze between one and five rare variants simultaneously in the pedigree sample used in the simulation study, we reached greater power by testing tens of variants together in annotated regions, or event hundreds of variants in the absence of annotations. Although the test is well-calibrated and powerful for extended pedigrees, we have demonstrated that it performs well when applied to small family structures, with modest Type I error rate inflation (Figure S9). It is noteworthy that the simulated variant ORs did not depend on the variant MAF due to limitations of the simulation software. The MAF-dependent variant weighting scheme of RetroFun-RVS was thus misspecified in the power evaluation. Greater power gains of RetroFun-RVS over the competing methods ignoring variant MAF could have been achieved had the variant ORs be inversely related to MAF. Some analyses have shown that Type I error rate or power are highly dependent on the number of variants present in the region of interest. Indeed, we have observed that when large numbers of variants are considered, RetroFun-RVS might provide conservative results involving some power loss (Figure S3), while a small number of variants tends to offer inflated Type I error rate (Figure S8). Complementary analyses are needed to inspect the empirical relationship between size of region and performance. Therefore, in the meantime we recommend in practice to use the covariance-adjusted model. Finally, bootstrap procedures (Figure S11) might be considered to sharply control type I error rate for small numbers of variants at the expense of longer computing time.

We argue that the performance of the proposed method is strongly dependent to the availability of the relevant tissue for the studied disease. Indeed, regulatory mechanisms operate in a tissue- or cell-type-specific manner. Our framework, by allowing the incorporation of several functional annotations from diverse tissues or cell-types without loss of power, is useful to highlight the underlying biological mechanisms involved in the trait. This aspect is central from a fine-mapping perspective, thus RetroFun-RVS will be an important tool to pinpoint causal variants located within non-coding regions, which could have been missed so far.

7 Data Availability

Cis-Regulatory Hubs and Topologically associated domains used in this paper are available on https://github.com/lmangnier/CRHs. Variant data were available from the 1000 Genome project: https://www.internationalgenome.org/data-portal/data-collection/phase-3. We have implemented RetroFun-RVS in a R package, available on GitHub (https://github.com/lmangnier/RetroFun-RVS).

8 Funding

Some data analyses were performed on computing resources from Compute Canada. This work was partly funded by the Canadian Statistical Sciences Institute through a Collaborative Research Team grant and by a National Science and Engineering Research Council of Canada Discovery grant to A Bureau.

9 Conflict of Interest

The authors declare that they have no conflict of interest.

References

Bureau, A., Begum, F., Taub, M. A., Hetmanski, J. B., Parker, M. M., Albacha-Hejazi, H., Scott, A. F., Murray, J. C., Marazita, M. L., Bailey-Wilson, J. E., Beaty, T. H., & Ruczinski, I. (2019). Inferring disease risk genes from sequencing data in multiplex pedigrees through sharing of rare variants. *Genetic Epidemiology*, *43*(1), 37–49. https://doi.org/10.1002/gepi.22155

Bureau, A., Parker, M. M., Ruczinski, I., Taub, M. A., Marazita, M. L., Murray, J. C., Mangold, E., Noethen, M. M., Ludwig, K. U., Hetmanski, J. B., Bailey-Wilson, J. E., Cropp, C. D., Li, Q., Szymczak, S., Albacha-Hejazi, H., Alqosayer, K., Leigh Field, L., Wu-Chou, Y. H., Doheny, K. F., ... Beaty, T. H. (2014). Whole Exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics*, *197*(3), 1039–1044. https://doi.org/10.1534/genetics.114.165225

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, *51*(12), 1664–1669. https://doi.org/10.1038/s41588-019-0538-0

Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K., Nasser, J., Jagadeesh, K., Weiner, D., Shi, H., Fulco, C., O'Connor, L., Pasaniuc, B., Engreitz, J. M., & Price, A. L. (2022). Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. *Nature Genetics*, 1546–1718. https://doi.org/10.1038/s41588-022-01087-y

He, Z., Xu, B., Lee, S., & Ionita-Laza, I. (2017). Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *American Journal of Human Genetics*, *101*(3), 340–352. https://doi.org/10.1016/j.ajhg.2017.07.011

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., & Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*, *7*(2). https://doi.org/10.1371/journal.pgen.1001289

Laird, N. M., & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, *7*(5), 385–394. https://doi.org/10.1038/nrg1839

Li, B., Wang, G. T., & Leal, S. M. (2015). Generation of sequence-based data for pedigree-segregating Mendelian or Complex traits. *Bioinformatics*, *31*(22), 3706–3708. https://doi.org/10.1093/bioinformatics/btv412

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics*, *83*(3), 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024

Li, M., Boehnke, M., & Abecasis, G. R. (2006). Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *American Journal of Human Genetics*, *78*(5), 778–792. https://doi.org/10.1086/503711

Li, W., Baumbach, J., Mohammadnejad, A., Brasch-Andersen, C., Vandin, F., Korbel, J. O., & Tan, Q. (2019). Enriched power of disease-concordant twin-case-only design in detecting interactions in genome-wide association studies. *European Journal of Human Genetics*, *27*(4), 631–636. https://doi.org/10.1038/s41431-018-0320-2

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American Journal of Human Genetics*, *104*(3), 410–421. https://doi.org/10.1016/j.ajhg.2019.01.002

Ma, S., Dalgleish, J., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J. D., Chung, W. K., Aschard, H., Silverman, E. K., Cho, M. H., He, Z., & Ionita-Laza, I. (2021). Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(47), 1–12. https://doi.org/10.1073/pnas.2105191118

Ma, Y., & Wei, P. (2019). FunspU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing

data. *PLoS Genetics*, *15*(4), 1–21. https://doi.org/10.1371/journal.pgen.1008081

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, *5*(2). https://doi.org/10.1371/journal.pgen.1000384

Mangnier, L., Joly-Beauparlant, C., Droit, A., Steve, B., & Bureau, A. (2022). Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia. *Life Science Alliance*, *5*(5), 1–12. https://doi.org/10.26508/lsa.202101156

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. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., ... Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, *593*(7858), 238–243. https://doi.org/10.1038/s41586-021-03446-x

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, *7*(3). https://doi.org/10.1371/journal.pgen.1001322

Ott, J., Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, *12*(7), 465–474. https://doi.org/10.1038/nrg2989

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, *19*(8), 491–504. https://doi.org/10.1038/s41576-018-0016-z

Schaid, D. J., McDonnell, S. K., Riska, S. M., Carlson, E. E., & Thibodeau, S. N. (2010). Estimation of genotype relative risks from pedigree data by retrospective likelihoods. *Genetic Epidemiology*, *34*(4), 287–298. https://doi.org/10.1002/gepi.20460

Schaid, D. J., Mcdonnell, S. K., Sinnwell, J. P., & Thibodeau, S. N. (2013). Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. *Genetic Epidemiology*, *37*(5), 409–418. https://doi.org/10.1002/gepi.21727

Sherman, T., Fu, J., Scharpf, R. B., Bureau, A., & Ruczinski, I. (2019). Detection of rare disease variants in extended pedigrees using RVS. *Bioinformatics*, *35*(14), 2509–2511. https://doi.org/10.1093/bioinformatics/bty976

Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J. D., Bass, N., Bigdeli, T. B., Breen, G., Bromet, E. J., Buckley, P. F., Bunney, W. E., Bybjerg-Grauholm, J., Byerley, W. F., Chapman, S. B., Chen, W. J., Churchhouse, C., Craddock, N., Cusick, C. M., ... Daly, M. J.

(2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*, *604*(7906), 509–516. https://doi.org/10.1038/s41586-022-04556-w

Sul, J. H., Cade, B. E., Cho, M. H., Qiao, D., Silverman, E. K., Redline, S., & Sunyaev, S. (2016). Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. *American Journal of Human Genetics*, *99*(4), 846–859. https://doi.org/10.1016/j.ajhg.2016.08.015

Sun, B. B., Kurki, M. I., Foley, C. N., Mechakra, A., Chen, C. Y., Marshall, E., Wilk, J. B., Sun, B. B., Ghen, C. Y., Marshall, E., Wilk, J. B., Runz, H., Chahine, M., Chevalier, P., Christé, G., Kurki, M. I., Palotie, A., Daly, M. J., Palotie, A., … Runz, H. (2022). Genetic associations of protein-coding variants in human disease. *Nature*, *603*(7899), 95–102. https://doi.org/10.1038/s41586-022-04394-w

Wu, C., & Pan, W. (2018). Integration of enhancer-promoter interactions with GWAS summary results identifies novel schizophrenia-associated genes and pathways. *Genetics*, *209*(3), 699–709. https://doi.org/10.1534/genetics.118.300805

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Yao, L., Liang, J., Ozer, A., Leung, A. K.-Y., Lis, J. T., & Yu, H. (2022). A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01211-7

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, *24*(R1), R102–R110. https://doi.org/10.1093/hmg/ddv259

Zhao, L., He, Z., Zhang, D., Wang, G. T., Renton, A. E., Vardarajan, B. N., Nothnagel, M., Goate, A. M., Mayeux, R., & Leal, S. M. (2019). A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data. *American Journal of Human Genetics*, *105*(4), 822–835. https://doi.org/10.1016/j.ajhg.2019.09.006