BIOINFORMATICS

Vol. 00 no. 00 2013
Pages 1–6

# Characterization of complex genetic disease using exomic SNVs and gene expression data

Aziz M. Mezlini [1,2], Lalla M. Mouatadid [1] and Anna Goldenberg [1,2*]

[1]Department of Computer Science, University of Toronto, Canada.
[2]Sickkids hospital, Toronto, Canada.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

Despite the large number of published studies investigating complex genetic diseases, the mechanisms of the diseases and their causes remain elusive for the majority of the patients. If the goal of understanding the disease and designing personalised treatment strategies is ever to be reached, we need novel innovative approaches that take into account the biology of the problem in order to succeed, where GWAS and similar statistical analysis methods were underpowered. In this paper, we present a method for combining Exome sequencing data and gene expression data in order to aggregate the genetic variants and propose a disease mechanism that is common across the patients. Through an exploration of the protein-protein interaction network, our method identify clusters where the occurrences of either harmful genetic variants or genes abnormal levels of expression is frequent. These clusters characterize the common biological mechanism of the disease and allow a restricted number of candidates for patient by patient retrieval of the causal genetic variants.

**Contact:** anna.goldenberg@utoronto.ca

## 1 INTRODUCTION

Over the last few years, hundreds of published studies have focused on testing individual variants in order to better understand complex diseases. Methods such as GWAS test the association of individual common SNPs with a genetic disorder based on cases versus controls statistical tests. Such methods did not really lead to a good characterization of the disease because of their limited perspective (individual SNPs, and only the common ones) and their low statistical power especially for low allele frequency variants.

To overcome the weaknesses of GWAS-type studies, several improvements have been proposed. For example, GWAS over pairs of SNPs instead of individual variants testing [23] or hypothesis driven GWAS to mitigate the multiple hypothesis testing problem and narrow down the investigation to functionally interesting variants [19]. However, even with these improvements over GWAS, the results are still limited and our understanding of the disease is as best tentative and incomplete [10]. In general, the combination of the results of all GWAS studies investigating a particular genetic disease can explain less than 10% of the genetic component of that disease. [9].

*to whom correspondence should be addressed

Other studies are focused on the gene level, aggregating variants (usually rare SNVs) over genes and testing the SNVs counts, with many possible approaches such as region-based analysis of variants of intermediate and low frequency test (GRANVIL [11], variable threshold method [17]). Several recent publications even take a step further and focus on the pathway level to study whole pathways enrichment [6]. Not only are these SNPs-counts-type studies based on discussable hypotheses (SNPs counts are higher in cases than in controls), they are also limited in the sense that they can only investigate variants that are localized within (or around) genes. This is very unfortunate since the majority of genetic variants discovered are in intergenic regions and even the variants found to be associated with disease by previous studies (GWAS) are often intergenic variants (38% in [1]). The ENCODE project very recently underlined the importance of intergenic elements such as enhancers, silencers and promoters in gene regulation and disease association [3]. Other recent publications are highlighting the role of these intergenic regions in cancer [20]. Therefore, by ignoring genetic variants that are outside of genes, we risk passing by the real causes of genetic diseases. Moreover, most of these counting-based studies do not take into account additional information we may have on genetic variants such as harmfulness predictions and functional annotations.

In this paper, we present a new method to identify a disease mechanism that is a common factor among patients. Our approach not only uses Exome data and harmfulness predictions of variants within genes, but also uses gene expression data as a natural proxy for intergenic variants. Our hypothesis is that intergenic variants can only affect the disease through a change in the level of expression of proteins, unlike the Exomic variants which can affect the protein sequence itself. Moreover, it is important to note that gene expression data can also be considered a proxy for other potentially causal genetic and epigenetic variants such as CNVs, histone modification, methylation and structural variants which mostly affect levels of expression rather than the protein code.

Our approach combines both the exome sequences and genes expression in order to assess a gene possible role in the disease and explore the protein-protein interaction network to determine the mechanism of the disease.

By characterizing the exact mechanism that, when disrupted, is causing the disease, we will be able to accurately characterize the disease and narrow down its potential causes in most of the patients. We hope this approach will finally allow a large scale understanding of the disease causes and enable the development of patient-specific treatments.

# 2 METHODS

## 2.1 Overview of the algorithm

In this section, we present our approach for combining Exome data and expression data and for determining the disease mechanism. To combine the variants from Exome data (SNPs and Indels) with the quantitative observations from expression data, the first step is to transform both types of data into harmfulness predictions that are similar to probabilities estimations. The harmfulness prediction for a SNP or an indel is the estimation of the probability that the given SNP/indel disrupts the function of gene copy it resides on. Similarly, harmfulness predictions from gene expression can be thought of as estimations of probability that the level of expression of a given gene in a given patient is unusual enough to disrupt the gene function. In both cases, a harmfulness prediction can also be loosely assimilated to the expected lack (proportion) of functioning proteins formed from the considered gene.

After estimating harmfulness predictions for the Exome variants, aggregating them by genes (Section 2.2) and converting gene expression to harmfulness (Section 2.3), we can combine these predictions (Section 2.4) to obtain one value of harmfulness for every gene and for every individual.

Finally, we use a protein-protein interaction network in order to select subsets of genes that give the best patients-versus-controls separation (Section 2.5) when their harmfulness predictions are combined together. We use a greedy algorithm to construct possible disease mechanism while exploring the protein-protein interaction network (Section 2.6).

## 2.2 Aggregating exomic harmfulness predictions

We will interpret harmfulness predictions as the expected proportion of the gene expression that will be dysfunctional at the protein stage because of exomic SNVs. In this context, harmfulness predictions on each haplome should be computed and then the resulting harmfulness predictions of the haplomes will be combined.

$$H_{gene} = \frac{\min(1, \sum_{v \in M} H_v) + \min(1, \sum_{v \in P} H_v)}{2} \quad (1)$$

Where $H$ is a harmfulness prediction, $v$ is a variant (exomic SNV or indel), and $M$, $P$ are the two haplomes for the considered gene. The min function guarantees that the maximal harmfulness that can be done to a haplome is equivalent to having all proteins produced from that haplome dysfunctional (theoretically half the product of the gene unless we have evidence of differential allelic expression). Also, since an SNV $v$ is a polymorphism on a haplome, $H_v$ is always between 0 and 1. A homozygous SNV is simply counted as two variants with the same harmfulness prediction.

If complete phase information is unavailable, we will average our harmfulness prediction for the gene over all possible configurations of variants across the haplomes. Knowledge of which variants are or are not on the same haplome can be obtained from looking at the reads or from sequencing other family members. That knowledge can easily be included as additional constraints by not considering impossible configurations. A simple example is the case of two SNVs that constitute a homozygous variant: they are necessarily on different haplomes. Therefore any prediction on a configuration where the two SNVs are on the same haplome will not be contribute to the average.

To avoid extreme cases where the number of exomic SNVs in a gene and the number of possible configurations over haplomes is very large, we will use the following sampling algorithm:

1. Assign homozygous variants to both copies of the gene.
2. Randomly assign all other SNVs to the two copies.
3. Check if the configuration conform to any additional constraints.
4. If the step 3 is valid, compute the harmfulness as in equation 1 and store the result.
5. Repeat all previous steps until we have a certain number of stored values.
6. The gene harmfulness prediction is the average over the stored values.

## 2.3 Computing expression-based harmfulness predictions

Gene expression is easily affected by various factors. In particular, we should be aware that many genes will behave differently in the presence of the disease because of the natural reaction of the organism or because of the effects of drugs. Therefore, the naive approach that associate high harmfulness predictions for the most differentially expressed genes in respect to the healthy controls is bound to fail since in most diseases or infections hundreds of genes have very different levels of expression between affected and healthy individuals.

In our method, if the controls and most of the patients share a similar gene expression distribution, and a small subset of patients show a significantly different behaviour , we can assume that those particular patients have unusual (and therefore harmful) levels of expression. If however, the expression levels difference is sufficiently great between cases and controls, then we are probably observing an effect of either immune response or treatment. In that case, we can only assign high harmfulness predictions to the extremely unusual measurements and not to all the measurements that are simply different from controls.

## 2.4 Combining harmfulness predictions from different sources

We consider multiple sources of harmfulness such as gene expression harmfulness and SNPs harmfulness. The idea is to combine these harmfulness scores in order to assign a final harmfulness score to each gene, taking into consideration two aspects of the gene; its quality captured by the SNPs harmfulness prediction, and the quantity captured in the computed harmfulness of the expression. If we suppose x and y are two such harmfulness predictions, their combination is computed as follow:

$$Harm_{gene} = 1 - (1 - x)(1 - y) = x + y - xy \quad (2)$$

Where $Harm_{gene}$ reflects the harmfulness score of the gene. We can easily verify that the gene harmfulness defined here is still consistent with the definition of harmfulness predictions (probability between 0 and 1) and that the combined harmfulness is always higher than both primary predictions.

Similarly, if we consider the damage/harmfulness done to mechanism $M$ as the combination of the harmfulness predictions of the genes constituting the mechanism then the mechanism fitness is also the product of the fitnesses. In other words:

$$Harm_M = 1 - \prod_{Gene_i \in M} (1 - Harm_{Gene_i}) \quad (3)$$

## 2.5 Assigning scores to genes and mechanisms

For every gene we have two sets of harmfulness predictions: One set representing the patients and one for the controls. By comparing the distribution of harmfulness predictions in the patients to the distribution in the controls, we are able to give a score to genes that are likely to be associated with the disease and this will help us prioritize which genes to include in the disease mechanism.

We do not expect a majority of the patients to have in common one single affected gene. Instead we expect a gene which is part of the disease mechanism to be affected in at least a small proportion of the patients. The rest of the patients will be similar to the controls for the considered gene. Therefore, testing the difference between the mean of the patients distribution and the mean of the controls distribution is not the appropriate way of scoring the genes. Instead, to characterize the difference between the two distributions, we will focus on the forward tail that will better describe a subgroup of patients having abnormally high harmfulness predictions.

Let us characterize the tails of the distributions by considering only the individuals that have higher harmfulness than the 0.9 quantile of the controls' harmfulness predictions distribution for the considered gene. All other

harmfulness predictions are put to zero for both patients and controls. Then we look at the difference between the gene harmfulness predictions averages in cases and controls. If it is significantly larger than 0, the gene is then considered a good candidate to be included in the disease mechanism. Another alternative approach we used is to characterize the cases and controls distributions with a Kolmokorov-Smirnov test.

The scoring function is similar for a putative mechanism. The mechanism harmfulness prediction being the combination of the predictions for the genes constituting the mechanism (See Section 2.4 ), we will simply compare the combination's distribution over the cases and over the controls by focusing on the forward tails similarly to our approach with genes.

### 2.6 Constructing the disease mechanism

In this part, we will use the protein-protein interaction network to build a biologically meaningful mechanism of the disease.

The disease mechanism is first initialized to a candidate gene which have a good score according to Section 2.5. We will look at all the neighbours of order less than $k$ (the set of genes/nodes within k edges from the considered gene). For each one of these neighbours we will estimate the score of the mechanism constituted of that gene plus the initial gene. The neighbour which gives the best score will be added to the mechanism. We will iteratively look at the neighbours of order $\leq k$ of the mechanism in construction and select the best gene to be added until score of the mechanism no longer shows a substantial increase (for now we use a pre-determinate threshold). To avoid overfitting, we also add a penalty on the number of genes in a mechanism.

This whole process is repeated starting from different candidate genes in order to thoroughly explore the network. The resulting mechanisms are compared by their final scores and the best can be investigated for functional relevance.

## 3 RESULTS

### 3.1 Simulated data

We used the genes and interactions present in BioGrid human protein-protein interaction network.

For simulating the SNPs, we used the European population model with the optimal parameters in [7]. We generated the missense SNPs of 900,000 individuals with their selection coefficient. Overall there was around half a million SNPs most of which are very rare since the corresponding mutations happened in the recent population expansion period.

By considering the SNPs with selection coefficient greater than s=0.001 as harmful, we simulated Polyphen2 scores for all SNPs. For now, these scores are sampled from Gaussian with mean respectively equal 0.2 and 0.8 for non-damaging and damaging SNPs (The standard deviations are 0.3).

As for the expression data, we started from a healthy expression dataset (TODO: cite GEO ressource) and for every gene, we computed the observed distribution of expression measurements to be used as a background signal of expression. We simulated a number of perturbations of expression uniformly across genes and individuals.

Then we selected a number of genes that are close to each other on the protein-protein interaction network as being the causal disease mechanism. For every individual, we counted the number of harmful mutations and the number of expression perturbations within the disease mechanisms and we considered that sum as the phenotype. The cases were then selected to be the top 1000 patients (we suppose the disease prevalence in the population is 0.1%) and the controls were randomly sampled from the rest of the population.

For every individual selected (patient or control), we changed the expression levels according to the gene expression perturbations

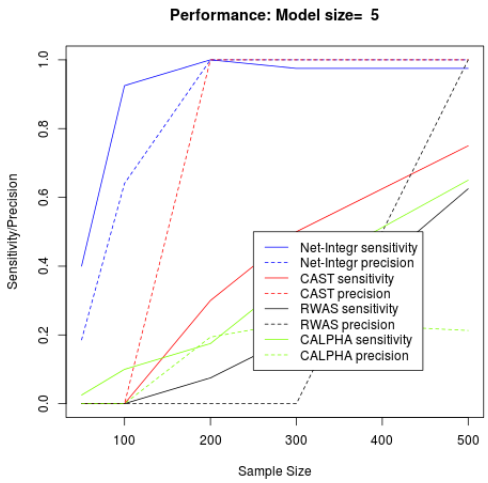


**Fig. 1.** Precision and recall as a function of the sample size used and comparison with the results of three variants aggregation methods. The simulated model always contain 5 genes and we fixed Expression-SNPS causality to 50% (Both have equal contribution to the phenotype on average). This curve is obtained by averaging over the results of 20 simulations

simulated earlier on. The magnitude of the change in expression was uniformly sampled from $[0, 4\sigma]$ where $\sigma$ is the standard deviation for the considered gene expression (background signal).

It is by design that we simulate any aberration as an event that cannot easily be individually separated from the noise, whether it is in SNPs (noisy Polyphen2 scores) or in expression (noisy perturbations). The goal of our method is to aggregate weak signals from all variants and all patients in order to detect the disease mechanism and that is the ability the simulations should test.

We varied our simulation parameters controlling the disease signal:

- Number of causal genes.
- Sample Size.
- Proportion of aberrations related to expression (versus SNVs).

We observe the effect of changing these parameters on the performance of our method. We first characterize the performance by how well was the disease mechanism reconstructed (precision, recall) and compared our results to those of three rare variants aggregation methods: CAST [5], C-ALPHA[12] and RWAS[18] (See Figures 1 and 2 ). The precision and recall for the other methods were measured by considering any gene with a p-value lower than the significance threshold as a positive ($< 8.10^{-6}$ after multiple hypothesis correction). We also estimate how often we can determine the exact causal aberrations (expression change or a particular SNP) in all patients after we narrowed down the possible causes to the genes in the disease mechanism.

Figures 1, 2 and 3 show that our method is much more powerful than any other one considered and that it is often the only method able to detect the disease mechanism in many realistic scenarios such as a relatively large number of causal genes involved or a large effect of gene expression aberrations or even a low sample size.

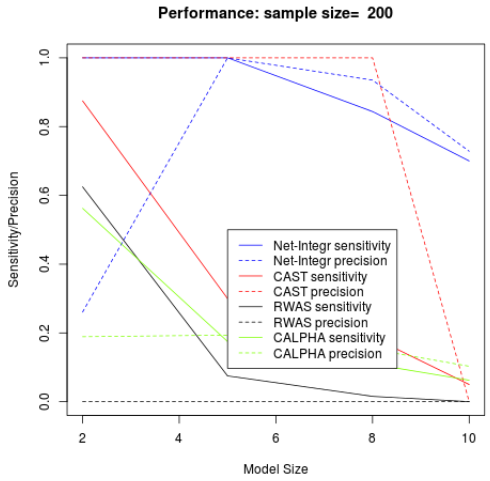


**Fig. 2.** Precision and recall as a function of the model size used (number of genes) and comparison with the results of three variants aggregation methods. We use a fixed sample size of 200 cases and 200 controls and a fixed Expression-SNPS causality of 50% (Both have equal contribution to the phenotype on average). This curve is obtained by averaging over the results of 20 simulations

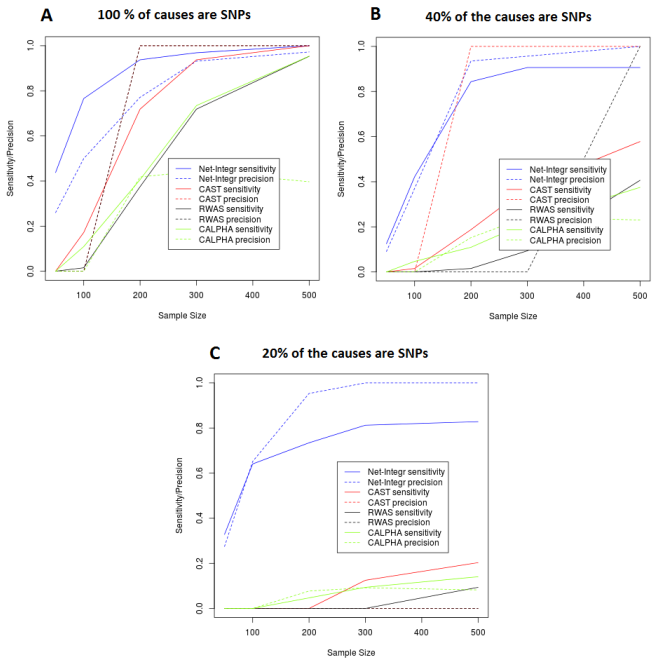


**Fig. 3.** Precision and recall as a function of the sample size used for different proportions of expression versus SNPs causality (**A** 100%, **B** 40%, **C** 20%). The simulated model always contain 8 genes. This curve is obtained by averaging over the results of 20 simulations

In particular, Figure 3 show the effect of varying the contribution of expression aberrations to the phenotype: There is a significant drop in power for all methods when expression is responsible for

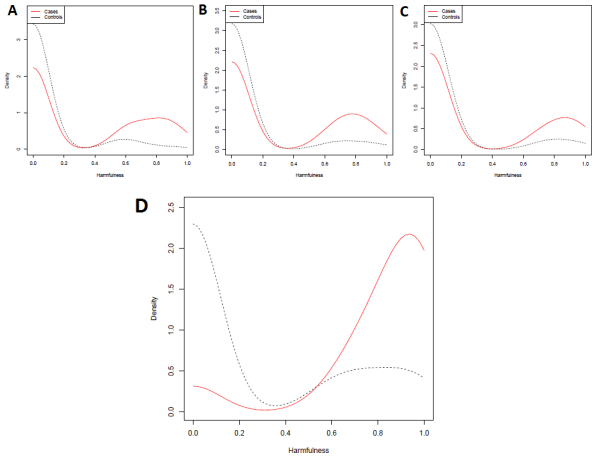


**Fig. 4.** Aggregating variants over the genes (A,B,C) of a mechanism greatly strengthen the disease mechanism signal (D) in simulations.

60% or 80% of the phenotype compared to when the phenotype is only the result of the SNPs within genes. However, our approach is the only one that remains able to detect the causal disease mechanism. Figure 3-C for example, is considering the case where 80% of the phenotype in the cases is related to expression aberrations in the involved genes, i.e regulatory variants (intergenic variants, epigenetic, etc) rather than protein coding variant. This scenario is not unlikely based on the proportion of variants associated with different diseases that are found to be outside versus within genes' coding regions.

Such high performance in characterizing the disease would not be possible if we considered individual genes only. An example of the importance of signal aggregation is shown in figure 4. The distribution of harmfulness predictions for genes A, B, C contains slightly more high values for some of the patients, than for controls. By combining the harmfulness of all 3 genes, we observe a much stronger signal in cases versus controls.

In all our current simulations experiments, we were able to successfully recover all the variants simulated as causal after we identify the correct disease mechanism. More in-depth analysis of the effects of the simulations parameters is in progress.

## 3.2 JIA

We perform additional experiments on real data using Juvenile Idiopathic Arthritis expression data and ImmunoChip data (instead of Exome data) on 160 patients classified into 7 different phenotypic subtypes.

As a first investigation of the data and because this dataset did not come with controls, we simulated our own ImmunoChip data controls from the minor allele frequencies (MAF) of the SNPs investigated by the ImmunoChip. We used dbSNP database to retrieve this information.

We first used snpNexus [4, 21, 22] which annotate all the SNPs by assigning them to genes and categorizing them into intronic, intergenic, UTR, non-sense, synonymous or non-synonymous. We assigned a harmfulness prediction of 1 to non-sense SNPs and to intronic SNPs on the splicing site. Tools such as Polyphen2 [2]

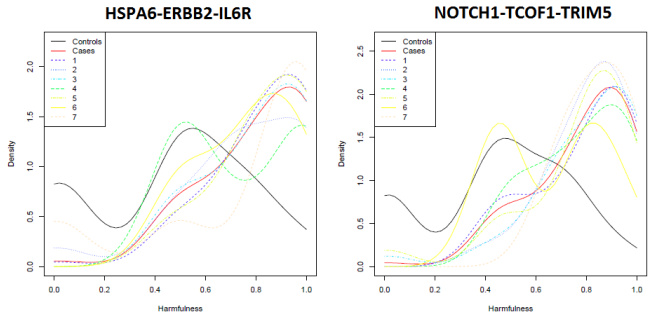


**Fig. 5.** Preliminary results on JIA data. The curves are the distributions of the mechanism harmfulness for cases and controls and across the seven clinical subtypes.

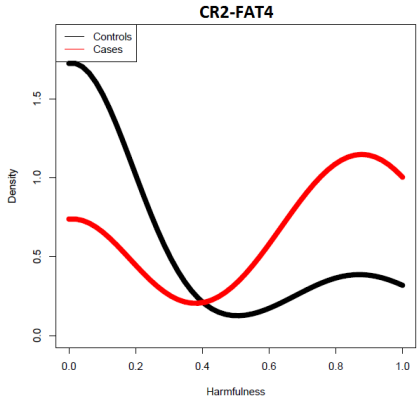


**Fig. 6.** Preliminary results on IBD data. The curves are the distributions of the mechanism harmfulness for cases and controls.

and SIFT [8, 16, 15, 14, 13] assign harmfulness predictions to the non-synonymous SNPs.

We run our method on this data (ImmunoChip only for now) and it showed promising signal on a few potential mechanisms (Fig 5) despite the fact that it is not the appropriate type of data for our method (not Exome, no controls...).

### 3.3 IBD

The Inflammatory Bowel disease dataset was downloaded from GEO website (Accession number GSE41269). It contains 149 Ulcerative Colitis patients (UC is a form of IBD) and 20 controls affected with Familial adenomatous polyposis (FAP). We have both SNP data (Illumina Omniexpress BeadChip :730000 SNPs) and gene expression data (Affymetrix Gene Expression Array : around 15000 genes measured) for all individuals.

After pre-processing the data and computing Polyphen2 scores similarly to what we did in JIA data, we ran our method on this dataset. Because of the larger sample size than in the JIA data and because of the presence of controls we were able to compute significance of our results by performing multiple labels permutations.

We identified the CR2 gene and the CR2-FAT4 gene pair as two potential mechanisms of the disease ($p = 0.03$ and $0.07$ respectively). Figure 6 shows the distribution of mechanism damage scores for patients and controls and it clearly shows a separation between the two groups.

In the future, we also plan to obtain results on real data through cross validation. We apply our model on a training set to obtain a disease mechanism and we test if the disease mechanism as a classifier model can separate cases from controls on an independent test set.

## 4 DISCUSSION

In this paper we present a new approach for aggregating genetic variants from exome and expression data in order to identify a disease mechanism that is common across patients. We demonstrated that solving the problem on typical data size is perfectly feasible on simulations where we are always able to retrieve the simulated disease mechanism. We started analysing real data and we already obtained promising preliminary results.

The most dangerous pitfall of our approach is the noisy nature of expression data: On one hand, it is true the effect of most of the noise (wet lab artefacts, measurements errors, variance between individuals) is reduced by our patients versus controls approach. On the other hand, gene expression data (especially if it is from blood) can also reflects the consequences of the disease and even the effects of used dugs. We need to use approaches that control for this. For example, if the expression data is from blood, we can control for the difference in populations of cells in the blood by using clinical data (leukocyte counts). This approach can correct for the natural immune response effect on expression data that could otherwise be mistaken for harmful levels of expression.

We also need to make sure that when combining expression data with Exome data we do not lose the disease signal into the noise. For example, we only combine them if it improve the overall score of the gene or if there is already a signal in the Exome data.

## ACKNOWLEDGEMENT

*Funding*:

## REFERENCES

[1]A Adeyemo and C Rotimi. Genetic variants associated with complex human diseases show wide variation across multiple populations. *Public health genomics*, 13(2):72–79, 2009.

[2]Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[3]Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[4]Claude Chelala, Arshad Khan, and Nicholas R Lemoine. Snpnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5):655–661, 2009.

[5]Thomas J Hoffmann, Nicholas J Marini, and John S Witte. Comprehensive approach to analyzing rare genetic variants. *PLoS One*, 5(11):e13584, 2010.

[6]Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[7]Gregory V Kryukov, Alexander Shpunt, John A Stamatoyannopoulos, and Shamil R Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871–3876, 2009.

[8]Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009.

[9]Dajiang J Liu and Suzanne M Leal. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *The American Journal of Human Genetics*, 2012.

[10]Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[11]Andrew P Morris and Eleftheria Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2):188–193, 2010.

[12]Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly. Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322, 2011.

[13]Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

[14]Pauline C Ng and Steven Henikoff. Accounting for human polymorphisms predicted to affect protein function. *Genome Research*, 12(3):436–446, 2002.

[15]Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[16]Pauline C Ng and Steven Henikoff. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.*, 7:61–80, 2006.

[17]Alkes L Price, Gregory V Kryukov, Paul IW de Bakker, Shaun M Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, 86(6):832, 2010.

[18]Jae Hoon Sul, Buhm Han, Dan He, and Eleazar Eskin. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. *Genetics*, 188(1):181–188, 2011.

[19]Lei Sun, Johanna M Rommens, Harriet Corvol, Weili Li, Xin Li, Theodore A Chiang, Fan Lin, Ruslan Dorfman, Pierre-François Busson, Rashmi V Parekh, et al. Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nature genetics*, 44(5):562–569, 2012.

[20]Qianzi Tang, Yiwen Chen, Clifford Meyer, Tim Geistlinger, Mathieu Lupien, Qian Wang, Tao Liu, Yong Zhang, Myles Brown, and Xiaole Shirley Liu. A comprehensive view of nuclear receptor cancer cistromes. *Cancer research*, 71(22):6940–6947, 2011.

[21]Abu Z Dayem Ullah, Nicholas R Lemoine, and Claude Chelala. Snpnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research*, 40(W1):W65–W70, 2012.

[22]Abu Z Dayem Ullah, Nicholas R Lemoine, and Claude Chelala. A practical guide for the functional annotation of genetic variations using snpnexus. *Briefings in bioinformatics*, 2013.

[23]Xuesen Wu, Hua Dong, Li Luo, Yun Zhu, Gang Peng, John D Reveille, and Momiao Xiong. A novel statistic for genome-wide interaction analysis. *PLoS genetics*, 6(9):e1001131, 2010.