

RESEARCH ARTICLE

# A Protein Domain and Family Based Approach to Rare Variant Association Analysis

Tom G. Richardson<sup>1</sup>, Hashem A. Shihab<sup>1</sup>, Manuel A. Rivas<sup>2</sup>, Mark I. McCarthy<sup>2,3</sup>, Colin Campbell<sup>4</sup>, Nicholas J. Timpson<sup>1</sup>, Tom R. Gaunt<sup>1\*</sup>

**1** MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, **2** Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom, **3** Oxford Centre for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford, United Kingdom, **4** Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom

\* [Tom.Gaunt@bristol.ac.uk](mailto:Tom.Gaunt@bristol.ac.uk)



## OPEN ACCESS

**Citation:** Richardson TG, Shihab HA, Rivas MA, McCarthy MI, Campbell C, Timpson NJ, et al. (2016) A Protein Domain and Family Based Approach to Rare Variant Association Analysis. PLoS ONE 11(4): e0153803. doi:10.1371/journal.pone.0153803

**Editor:** Patrick Lewis, UCL Institute of Neurology, UNITED KINGDOM

**Received:** September 28, 2015

**Accepted:** April 4, 2016

**Published:** April 29, 2016

**Copyright:** © 2016 Richardson et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** UK10K data is available from the European Genome-phenome Archive (<https://www.ebi.ac.uk/ega>) for researchers who meet the criteria for access to confidential data ([http://www.uk10k.org/data\\_access.html](http://www.uk10k.org/data_access.html)). The data are available to other researchers in the same way we received them. Data access is dependent upon a Data Access Agreement (DAA) agreed upon by UK10K and the requestors. Relevant data are from the UK10K COHORT ALSPAC (EGAS00001000090) and UK10K COHORT TWINSUK (EGAS00001000108) on EGA after obtaining a DAA with UK10K's Data Access Committee.

## Abstract

### Background

It has become common practice to analyse large scale sequencing data with statistical approaches based around the aggregation of rare variants within the same gene. We applied a novel approach to rare variant analysis by collapsing variants together using protein domain and family coordinates, regarded to be a more discrete definition of a biologically functional unit.

### Methods

Using Pfam definitions, we collapsed rare variants (Minor Allele Frequency  $\leq 1\%$ ) together in three different ways 1) variants within single genomic regions which map to individual protein domains 2) variants within two individual protein domain regions which are predicted to be responsible for a protein-protein interaction 3) all variants within combined regions from multiple genes responsible for coding the same protein domain (i.e. protein families). A conventional collapsing analysis using gene coordinates was also undertaken for comparison. We used UK10K sequence data and investigated associations between regions of variants and lipid traits using the sequence kernel association test (SKAT).

### Results

We observed no strong evidence of association between regions of variants based on Pfam domain definitions and lipid traits. Quantile-Quantile plots illustrated that the overall distributions of p-values from the protein domain analyses were comparable to that of a conventional gene-based approach. Deviations from this distribution suggested that collapsing by either protein domain or gene definitions may be favourable depending on the trait analysed.

**Funding:** Funding for UK10K was provided by the Wellcome Trust under award WT091310. This work was supported by the UK Medical Research Council (MRC IEU MC\_UU\_12013/8). TR is a UK MRC PhD Student.

**Competing Interests:** The authors have declared that no competing interests exist.

## Conclusion

We have collapsed rare variants together using protein domain and family coordinates to present an alternative approach over collapsing across conventionally used gene-based regions. Although no strong evidence of association was detected in these analyses, future studies may still find value in adopting these approaches to detect previously unidentified association signals.

## Introduction

Despite the success in identifying genetic associations with complex disease in recent years, we are still relatively unaware of the proportion of this phenotypic variation that rare variants are responsible for. Study designs over the last decade have largely concerned common variants, and whilst the amount of additive genetic variance explained by these variants is greater than initially expected[1], the case of the “missing heritability” still remains. Endeavors have therefore shifted to uncover the role of rare variants, with potentially much larger effect sizes than those observed from common variants[2, 3]. Due to breakthroughs in next generation sequencing we now have a wealth of data consisting of rare variants, paving the way for the development of novel methodology that allows us to investigate the impact of rare genetic variation on complex disease[4]. These methods should become particularly useful once next generation sequencing becomes more extensively undertaken in large population collections.

One type of approach involves grouping all variants within the same gene together, as they are likely to mark functional effects on the same protein or RNA, followed by analysing the combined effect of these variants using recently developed association tests. However, typically these approaches are underpowered[5]. A major cause of this is due to collapsing variants with contrasting directions of effect, as well as variants with little to no effect (neutral variants), which has inspired the development of variance-component tests (e.g. C-Alpha[6], SKAT[7]). Another plausible explanation for this could be that variants within the same genomic region that are grouped together do not necessarily share similar function. Analyses may therefore benefit from redefining the region of interest on the basis of likely functional consequences. Doing so may identify a more unified potential effect from sets of variants, not observed when collapsing across entire genes.

Protein domains are distinct functional, structural and evolutionary units which can either span sub-sections of a protein or its entire length. The exploration of human disease mutations have found that they tend to cluster together within highly conserved protein positions[8, 9], which certain protein domains occupy[10]. Amongst other functional tasks, protein domains can interact with each other and lead to protein-protein interactions (PPIs). It is understood that mutations that affect the binding interface of proteins can lead to dysfunctional allosteric changes which can have a downstream effect on disease[11]. Domains which consist of the same DNA sequence may occur multiple times across the genome in different genomic regions but can have a similar functional consequence. These regions are collectively referred to as protein families.

In comparison to conventional approaches which aggregate variants according to gene coordinates, we have undertaken a study to evaluate whether collapsing variants across regions based on protein domain coordinates provides a viable alternative. We aggregated variants together using 3 different definitions:

1. **Protein domains:** Individual regions of the genome which mapped to individual protein domains. Certain protein domains can consist of the entire protein, which would therefore

result in an identical analysis to using gene coordinates when analysing variants in exons. The exception to this would be variants at the 3' and 5' locations. However, protein domains can also consist of subsections of the protein, which would therefore only map back to a subsection of the corresponding gene region. These individual regions therefore varied in length based on definitions. In these instances we hypothesised that using domain coordinates may aggregate together more functionally relevant variants, thus resulting in a gain of statistical power in the resulting analysis.

2. **Domain-Domain Interactions:** As two individual protein domains can interact and subsequently lead to the formation of PPIs, it is plausible that the corresponding pair of genomic regions may harbour functionally relevant variants. Whilst a mutation in either domain may be sufficient to affect binding affinity, analysing these regions together may provide stronger evidence of association than analysing each individually as they affect the same interaction. Consequently this definition consists of all variants from within two of the regions analysed using the previous definition, which were predicted to be responsible for a PPI.
3. **Protein Families:** As previously mentioned, protein families can be defined as multiple protein domains which have a similar sequence and structure. This final definition therefore consists of multiple genomic regions which map to the same type of protein domain. Certain protein families can consist of a large frequency of domains and in these circumstances it seemed unlikely that so many different proteins would be involved along the causal pathway of disease. However, as these protein domains can have similar functionality it was worthwhile investigating whether variants within these regions, after selecting only genes with experimental evidence of interaction, were collectively associated with disease.

We hypothesised that, in comparison to conventionally used gene coordinates, aggregating rare variants together across these alternative definitions may result in a greater proportion of variants which have a similar impact along the causal pathway and fewer neutral variants which dilute the observed signal. The potential trade-off to this is that variants with a similar functional effect within the same gene region may end up being collapsed separately. Rare variant analyses using these alternative definitions were conducted using samples from individuals involved in the UK10K project, which consists of participants drawn from the ALSPAC (Avon Longitudinal Study of Parents and Children) and TwinsUK cohorts.

## Methods

### Cohort Description

The UK10K consortium has two main project arms. In this study, we have used data from the cohorts' arm which was designed to investigate the contribution of genome wide genetic variation to a range of quantitative traits. This arm contains individuals from two intensively studied cohorts of European ancestry, ALSPAC (Avon Longitudinal Study of Parents and Children) and TwinsUK:

**ALSPAC.** ALSPAC is a population-based cohort study investigating genetic and environmental factors that affect the health and development of children. The study methods are described in detail elsewhere[12, 13] (<http://www.bristol.ac.uk/alspac>).

Ethical approval was obtained from the National Research Ethics Service (NRES) Committee, South East London, REC 2. Written informed consent was obtained from parents for all measurements made.

**TwinsUK.** The TwinsUK registry is a cohort of volunteer adult twins from all over the United Kingdom[14]. Initially, only middle-aged women were recruited and as a result 83% of

the registry is female. The registry currently contains 51% monozygotic (MZ) and 49% dizygotic (DZ) twins aged 18–103 years. Further details are available online (<http://www.twinsuk.ac.uk/>).

Informed consent was obtained from participants before they entered the study and ethical approval was granted by the National Research Ethics Service (NRES) Committee, Westminster, London.

## Sequencing Data

DNA Samples from 4,030 UK10K study participants (2,040 offspring from the ALSPAC cohort, 1,990 from the TwinsUK cohort) were subjected to low coverage (6–8x average read depth) whole-genome sequencing (WGS). Sequencing was performed at both the Wellcome Trust Sanger Institute (WTSI) and the Beijing Genomics Institute (BGI). DNA (1–3μg) was sheared to 100–1000 bp using a Covaris E210 or LE220 (Covaris, Woburn, MA, USA). Sheared DNA was size subjected to Illumina paired-end DNA library preparation. Following size selection (300–500 bp insert size), DNA libraries were sequenced using the Illumina HiSeq platform as paired-end 100 base reads according to manufacturer's protocol.

Data that passed quality control (QC) was aligned to the GRCh37 human reference used in phase 1 of the 1000 Genomes Project. Reads were aligned using BWA (v0.5.9-r16)[15]. Of the 4,030 participants, 3,910 samples (1,976 ALSPAC and 1,934 TwinsUK) went through the variant calling procedure. Low quality samples were identified by comparing the samples to their GWAS genotypes using about 20,000 sites on chromosome 20. A total of 112 samples (48 ALSPAC and 64 TwinsUK) were removed, leaving 3,798 samples (1,928 ALSPAC and 1,870 TwinsUK) that were eligible for the genotype refinement phase.

Missing and low-confidence genotypes in the filtered VCFs were refined out using the imputation procedure in BEAGLE 4[16] with default parameters. Additional sample-level QC steps were carried out on refined genotypes, resulting in 17 samples (16 TwinsUK and 1 ALSPAC) being removed due to either non-reference discordance with GWAS SNV data >5%, multiple relations to other samples or failed sex check. A principal components analysis was conducted using EIGENSTRAT[17] to exclude participants of non-European ancestry after merging our data with a pruned 11 HapMap3 population dataset[18]. 44 subjects (12 TwinsUK and 32 ALSPAC) did not cluster to the European (CEU) cluster and were removed. The final sample size for association analyses comprised of 3,621 individuals which did not include any related pairs (1,754 TwinsUK and 1,867 ALSPAC).

## Data Collection

**UK10K Phenotypes.** **ALSPAC:** Non-fasting blood samples were taken from participants who attended the age 9 clinic (mean age: 9.9, range: 8.9–11.5). Plasma lipid concentrations (total cholesterol (TC), triglycerides (TG) and high density lipoprotein cholesterol (HDLc)) were measured by modification of the standard Lipid Research Clinics Protocol with enzymatic reagents for lipid determination[19]. Low density lipoprotein cholesterol (LDLc) concentration was subsequently calculated using the Friedwald equation[20]:

$$LDLc = TC - (HDLc + TG \times 0.45)$$

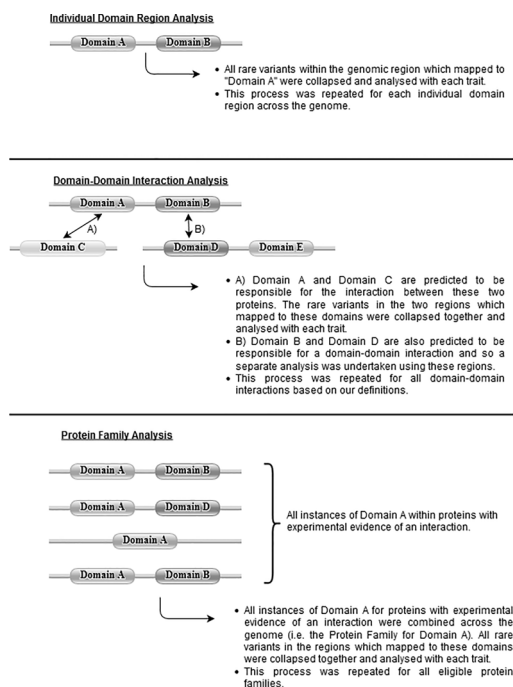
**TwinsUK:** Blood samples were taken after at least 6 hours of overnight fasting. The samples were immediately inverted three times and left to rest for 40 minutes at 4°C to obtain complete coagulation. The samples were then centrifuged for 10 min at 2000g and serum was removed. Four aliquots of 1.5 ml were placed into skirted micro centrifuge tubes and then stored in a -45°C

freezer until sampling[21]. A colorimetric enzymatic method was used to determine TC, TG and HDL-c levels. The Friedewald equation was used to calculate LDL-c levels in subjects.

## Statistical Analysis

**Simulated Data Analysis.** As a proof of concept, we first undertook analyses using simulated data consisting of 3,000 individuals each with 5,000 rare variants (i.e.  $MAF \leq 1\%$ ). 4,000 of these were neutral variants and 1,000 were disease causing variants (i.e. variants that were associated with the dichotomous phenotype). The odds ratios for the disease causing variants varied from 1.20 to 1.50, whereas the neutral variants only ever contributed to an odds ratio of 1.0. We randomly allocated variants into 250 ‘synthetic’ genes (i.e. not based on any genomic location, simply blocks of variants) of varying number of variants per gene (from 10 to 30 variants). Within these genes we defined subgroups of variants to resemble domains, varying in length between 5–20 variants. 1,000 of these synthetic domains were randomly selected for analysis. We conducted gene-based and domain-based analysis using the sequence kernel association test (SKAT)[7] to examine association between groups of variants within these regions and the synthetic phenotype. All simulated data was generated using PLINK v1.9[22].

**Application to real data.** Fig 1 summarises the planned analyses of this study. Using the UK10K sequence data, we took all Pfam protein family and domain coordinates and mapped them back to genomic coordinates using a custom Python script. These coordinates were extracted from the hg19 download used by the prediction tool Mutation Assessor [23].



**Fig 1. Graphical summary of how rare variants within domain regions were collapsed together for analysis.** This figure describes how regions of rare variants were collapsed together and analysed: • Individual Domains: Regions which mapped to individual domains of a protein were analysed as opposed to the conventional approach of using individual gene regions. • Domain-Domain Interactions: Pairwise regions which mapped to two domains predicted to be responsible for a domain-domain interaction were analysed together. • Protein Families: All regions which mapped to the same type of protein domain within proteins which had experimental evidence of interaction were combined and analysed together. Domain images were created using the generate graphics feature from Pfam (located at [http://pfam.xfam.org/generate\\_graphic](http://pfam.xfam.org/generate_graphic)).

doi:10.1371/journal.pone.0153803.g001

Variants were filtered to only include those with a  $MAF \leq 1\%$  and a CADD (Combined Annotation Dependent Depletion) C-Score  $\geq 15$ . This threshold is suggested by the authors of CADD as it equates to the 5% most deleterious variants across the genome as predicted by this resource.

We conducted analyses on regions based on three novel definitions of a functional unit. These were:

1. **Protein domains:** Individual genomic regions which mapped to individual protein domains according to Pfam definitions. These regions comprised of either sections of proteins or in some cases the whole protein. For the latter, analyses would have been identical to analysing variants within the entire exonic region of the gene with the exception of variants at the 3' and 5' locations.
2. **Domain-Domain Interactions:** iPfam was used to identify which pairwise domains were predicted to be responsible for a given protein-protein interaction according to STRINGdb v9.1 [24]. Firstly, all pairwise protein interactions which had at least some experimental evidence and with a STRING score  $\geq 0.8$  were extracted. Then for each pairwise interaction it was verified whether any domains from the first protein interacted with any domains in the second protein, according to iPfam. If this was true, the two domain regions in questions were added to the list of eligible domain-domain interactions (DDI) for our analysis. All variants within regions that were responsible for each domain-domain interaction were aggregated and analysed together (i.e. all variants within two of the regions in the previous analysis which were predicted to be involved in a PPI).
3. **Protein Families:** Variants were collapsed together across regions that were located within the same type of domain across the genome (i.e. variants within all regions which had the same Pfam ID). However, only domains within gene regions whose product had experimental evidence of interaction according to STRINGdb (again using a STRING score  $\geq 0.8$  threshold) were combined. This meant that variants within multiple regions involved in the initial analysis (i.e. the individual domain analysis) were analysed together here.

Only regions which had at least 2 remaining variants were analysed using SKAT with each lipid trait (HDL, LDL, TC and TG) in turn. All traits were inverse normal transformed prior to analysis. Further details on trait standardization can be found in the Supplementary Material (S1 File).

To evaluate whether results provided strong evidence of association we used a threshold for multiple comparisons using the Bonferroni correction (i.e.  $0.05/\text{number of regions analysed}$ ). All individual protein domains regions were reanalysed using the SKAT-O test [25] found to have more power than SKAT in situations where variants within a region have the same direction of effect [26]. A single variant analysis was also conducted using each lipid trait for all rare variants which were analysed previously. This was to ensure that aggregating variants together across regions was not causing evidence of association observed from a single variant analysis to consequently become undetected. These results were plotted using Quantile-Quantile (Q-Q) plots.

**Comparison between Individual Domain and Gene-based Results.** Q-Q plots were generated using the distribution of p-values from the results of the individual protein domain analysis with results from a conventional gene-based analysis. This analysis was undertaken with the same dataset but using gene start and end coordinates according to hg19 definitions and analysed as before using SKAT with each lipid trait. The quantiles from the domain-based analyses were therefore interpolated as there were more individual domains analysed



compared to the number of genes[27]. Q-Q plots were generated using the R package 'qqman'[28]. All statistical analyses were undertaken using R statistical software[29].

## Results

### Simulated Data analysis

Analysing groups of variants from within simulated domain regions provided stronger evidence of association in comparison to collapsing by gene regions, as 23 gene-based results survived the correction for multiple testing ( $P < 2.0 \times 10^{-4}$  (250 tests)) in comparison to 47 domain-based results ( $P < 5.0 \times 10^{-5}$  (1,000 tests)). The tops hits from these analyses can be found in [S1 File](#). This was due to analysing smaller blocks of variants which consisted of a higher proportion of disease causing variants (i.e. a smaller proportion of neutral variants were involved in these analyses and therefore incorporated less statistical noise). It was therefore hypothesised that, if a sufficient proportion of causal variants resided within domain regions, that our planned analyses should identify associational signals which would not be detected using a gene-based approach.

### Sample Characteristics

5,330,943 sites were excluded from further analyses due to showing Sanger/BGI batch effects, failed the test for Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-6}$ ) or were below the VQSLOD score cut-off (Variant Quality Score Recalibration) that corresponds to the maximum truth sensitivity tranche of 99.5% compared to HapMap3.3. Filtering to only include variants with a CADD C-Score  $\geq 15$  reduced the final number to 546,334.

After removing samples that failed QC, we were left with a sample size of 3,621 which did not include related pairs or non-European individuals. Subsequently removing individuals with missing phenotype information resulted in a final sample size of ~3,200 (3,210 for HDL, 3,191 for LDL, 3,206 for TC and 3,202 for TG).

### Individual Domain Analysis

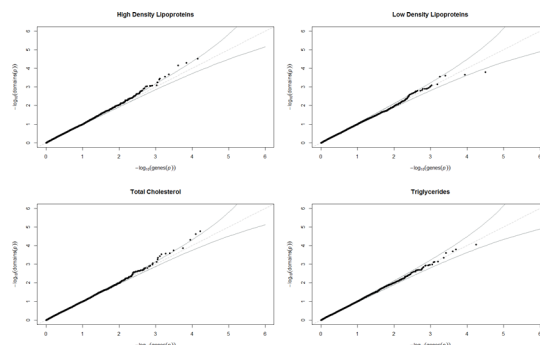
We mapped 76,391 protein families and domains back onto the human genome. 39,016 of which were canonical and used for all analyses. After filtering to only include variants with a CADD C-Score  $\geq 15$ , ~9,570 domains contained at least 2 variants after applying a MAF cutoff of 1%. These varied very slightly due to the small differences in sample size after matching on phenotypes. Our threshold for multiple comparisons were therefore  $5.22 \times 10^{-6}$ . No results from this analysis survived this threshold using either the SKAT or SKAT-O tests ([S1 File](#)).

### Domain-Domain Interaction Analysis

Using the iPfam and STRINGdb databases we predicted there to be 14,046 combined regions that were responsible for domain-domain interactions. After filtering to include SNVs with a CADD C-Score  $\geq 15$ , there were ~10,020 using the 1% MAF cutoff. This determined that our threshold for multiple comparisons was  $4.99 \times 10^{-6}$ . However, no results from the analysis survived this threshold ([S1 File](#)).

### Variants Collapsed by Protein Family Analysis

2,356 unique Pfam identifiers were used in the previous analyses. We collapsed all variants together within regions with the same Pfam identifier and then stratified them to only include SNPs with a CADD C-Score  $\geq 15$ . There were 3,114 regions with 2 or more variants in after filtering using the MAF cutoff of 1%, meaning our thresholds for multiple comparisons was



**Fig 2. Quantile-Quantile plots to compare distributions of p-values identified using individual domain and gene-based approaches to rare variant analysis.** These Q-Q plots represent the distribution of p-values from an analysis where rare variants have been collapsed together using protein domain coordinates and analysed with lipid traits. The reference distribution for these plots are distributions of p-values from an identical analysis except collapsing rare variants using conventional gene-based coordinates.

doi:10.1371/journal.pone.0153803.g002

$1.61 \times 10^{-5}$ . We analysed these regions as before and did not observe any p-values lower than the thresholds for multiple comparisons ([S1 File](#)).

Single variant analyses were undertaken for all variant analysed previously, to ensure that collapsing variants together did not result in any association signals becoming undetectable when using this approach. The results were plotted using Q-Q plots, which did not suggest that strong evidence of association was detected using single variant analyses for each lipid trait ([S1 File](#)).

## Comparison between Individual Domain and Gene-based Results

Q-Q plots which compare the distribution of p-values from the individual domain and gene-based analyses were generated. Results varied according to lipid trait, as the HDL and TC analyses suggested that the individual domain approach provided stronger evidence of association due to an uptick in signal. In contrast, the LDL and TG plots are predominantly confined within the 95% confidence intervals. These plots can be found in [Fig 2](#).

## Discussion

We have undertaken a novel approach to rare variant analysis which, to our knowledge, is the first of its kind. Detecting strong evidence of association from multiple rare variants over an entire gene region is challenging for many reasons, not least of all the possibility that these variants may reside in different types of structural and functional domains. By aggregating variants together across protein domains and families, we hypothesised that sets of variants may be more likely to have a similar functional impact, as well as contain fewer neutral variants, than when collapsing variants across entire genes. However, we have been unable to provide evidence to support this in our study. Future studies should therefore contemplate applying this approach to large-scale sequence data to further evaluate whether collapsing variants in this manner may identify association signals not detected using gene-based approaches.

It has become common practice for studies to undertake collapsing approaches using entire gene regions, although typically their findings have been underwhelming. This is in no small part due to a large proportion of neutral variants in collapsed regions which incorporate statistical noise into the analysis. Furthermore, regarding genes as functional units can have limitations. For instance, the protein product of a gene may contain multiple domains which can be recombined in a different order and alter its overall function, which consequently can cause a



different phenotypic effect downstream[30]. The study of protein domains has previously revealed important functional insights, such as the identification of *LRRK2* as a promising therapeutic target for the treatment of Parkinson's disease[31]. Moreover, there have been studies which have shown both Mendelian disease and somatic cancer mutations to cluster within certain types of protein domains[32, 33]. Proteins have evolved through the shuffling of functional domains, causing some domain sequences to be located many times across the proteome. Protein-protein interfaces are typically more conserved than the rest of the protein surface which is why in this study we examined the combined effect of rare variants that were responsible for domain regions predicted to interact with one and other.

A limitation to using protein domain regions when aggregating rare variants together is that the extent in which protein domains are categorised varies significantly. A subset of protein domains have been thoroughly investigated, whereas the functional role of the majority of domains remains unknown. Pfam definitions are based on amino acids sequences which are repeated across the proteome according to hidden Markov models, rather than any evidence implicating them in the aetiology of complex disease. However, as future research continues to develop our understanding of the functional task of protein domains, as well as our overall understanding of the genetic architecture of complex disease, there will be additional value in aggregating variants in the manner undertaken in this study.

In terms of how we can proceed by analysing rare variants in this manner, there are valuable resources that can aid domain-centric analyses. Along with the Pfam database used here, the Conserved Domains Database (CDD)[10], the SUPERFAMILY[34] database and the Protein Analysis Through Evolutionary Relationships (PANTER)[35] resource contain a wealth of information which can be utilised to aid rare variant association analyses. Furthermore, we have used STRINGdb in this study to define PPIs as it allowed us to filter only those interactions with experimental evidence and high confidence (i.e. a STRING Score of  $\geq 0.8$ ). However, resources such as IntAct[36] contain experimentally verified binding sites and could therefore be incorporated into the analysis pipeline presented in this study. A catalogue of mutation pathogenicity prediction tools have been developed in recent years to prioritise or weigh SNPs in association studies and many of these tools use conservation score as a key variable in their predictions. In this study we have used annotations from CADD, although prediction tools such as FATHMM-MKL[37] and DANN[38] may also be useful for variant filtering.

Despite a lack of statistically robust findings in our analyses, there may still be value in examining association signals from rare variants collapsed across highly conserved regions. Moreover, the parameters and resources we have used in this study for the analysis pipeline can be adjusted and this may lead to stronger evidence of association than observed here. For instance, the resources used to define our protein domain regions, the tool used to quantify the predicted deleterious impact of variants (as well as the threshold applied to filter) and the collapsing method used to analyse genotype-phenotype associations are all variables which can be adjusted for in future studies.

In this study, we have focused on lipid traits due to the success reported by other studies in recent years [39–41]. Although these traits are typically observed to be polygenic in nature [42], there are also monogenic diseases which can be caused by extreme lipid levels, such as familial hypercholesterolemia[43]. Mutations in genes such as *LDLR*, *APOB*, *PCSK9* and *LDLRAP1* are known to lead to this condition [44], although endeavours in rare variant analysis hope to underpin novel loci which harbour causal variants in lipid related diseases. However, it is also expected that rare variants may be causal to rarer diseases and thus the approach used in this study may be useful for future studies which wish to investigate this. Moreover, the studies which have had success in detecting novel loci in disease using rare variant approaches have used large sample sizes (i.e. Surakka et al had a sample size of over 60,000 individuals), in

comparison to the 3,200 individuals analysed in this study. Applying the approach outlined in this study to similarly large sample sizes may therefore yield improved results. The use of the software RAREMETAL[45] could be incorporated into the framework presented in this study to facilitate analyses using samples which include multiple cohorts.

Previous studies have suggested that larger sample sizes and alternative statistical methodology should help improve findings for collapsing methods. Using a more discrete definition of a functional unit across the genome, such as protein domains and families, provides a feasible alternative to collapsing by gene coordinates, which may yield biologically meaningful inferences and previously unidentified association signals when undertaking rare variant analyses.

## Supporting Information

**S1 File. Supporting Information.** Supplementary Material.  
(DOCX)

## Acknowledgments

This study makes use of data generated by the UK10K Consortium, derived from samples from the ALSPAC and TwinsUK data sets. A full list of the investigators who contributed to the generation of the data is available from [www.UK10K.org](http://www.UK10K.org). Funding for UK10K was provided by the Wellcome Trust under award WT091310.

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, manager, receptionists and nurses. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and Tom R. Gaunt will serve as guarantors for the contents of this paper. This work was supported by the UK Medical Research Council (MRC Integrative Epidemiology Unit, MC UU 12013/8). TGR is a UK MRC PhD student.

## Author Contributions

Conceived and designed the experiments: TR MM MR TG. Analyzed the data: TR. Wrote the paper: TR HS MM CC NT TG. Mapped domains coordinates: HS.

## References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860–921. doi: [10.1038/35057062](https://doi.org/10.1038/35057062) PMID: [11237011](https://pubmed.ncbi.nlm.nih.gov/11237011/).
2. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews Genetics*. 2010; 11(6):415–25. doi: [10.1038/nrg2779](https://doi.org/10.1038/nrg2779) PMID: [20479773](https://pubmed.ncbi.nlm.nih.gov/20479773/).
3. Gibson G. Rare and common variants: twenty arguments. *Nature reviews Genetics*. 2011; 13(2):135–45. doi: [10.1038/nrg3118](https://doi.org/10.1038/nrg3118) PMID: [22251874](https://pubmed.ncbi.nlm.nih.gov/22251874/).
4. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111(4):E455–64. doi: [10.1073/pnas.1322563111](https://doi.org/10.1073/pnas.1322563111) PMID: [24443550](https://pubmed.ncbi.nlm.nih.gov/24443550/); PubMed Central PMCID: PMC3910587.
5. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB. The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS genetics*. 2012; 8(2):e1002496. doi: [10.1371/journal.pgen.1002496](https://doi.org/10.1371/journal.pgen.1002496) PMID: [22319458](https://pubmed.ncbi.nlm.nih.gov/22319458/); PubMed Central PMCID: PMC3271058.

6. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. *PLoS genetics*. 2011; 7(3):e1001322. doi: [10.1371/journal.pgen.1001322](https://doi.org/10.1371/journal.pgen.1001322) PMID: [21408211](https://pubmed.ncbi.nlm.nih.gov/21408211/); PubMed Central PMCID: PMC3048375.
7. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*. 2011; 89(1):82–93. doi: [10.1016/j.ajhg.2011.05.029](https://doi.org/10.1016/j.ajhg.2011.05.029) PMID: [21737059](https://pubmed.ncbi.nlm.nih.gov/21737059/); PubMed Central PMCID: PMC3135811.
8. Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. *Human molecular genetics*. 2001; 10(21):2319–28. PMID: [11689479](https://pubmed.ncbi.nlm.nih.gov/11689479/).
9. Mooney SD, Klein TE. The functional importance of disease-associated mutation. *BMC bioinformatics*. 2002; 3:24. PMID: [12220483](https://pubmed.ncbi.nlm.nih.gov/12220483/); PubMed Central PMCID: PMC128831.
10. Fong JH, Marchler-Bauer A. Protein subfamily assignment using the Conserved Domain Database. *BMC research notes*. 2008; 1:114. doi: [10.1186/1756-0500-1-114](https://doi.org/10.1186/1756-0500-1-114) PMID: [19014584](https://pubmed.ncbi.nlm.nih.gov/19014584/); PubMed Central PMCID: PMC2632666.
11. Gonzalez MW, Kann MG. Chapter 4: Protein interactions and disease. *PLoS computational biology*. 2012; 8(12):e1002819. doi: [10.1371/journal.pcbi.1002819](https://doi.org/10.1371/journal.pcbi.1002819) PMID: [23300410](https://pubmed.ncbi.nlm.nih.gov/23300410/); PubMed Central PMCID: PMC3531279.
12. Golding J, Pembrey M, Jones R, Team AS. ALSPAC—the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatric and perinatal epidemiology*. 2001; 15(1):74–87. PMID: [11237119](https://pubmed.ncbi.nlm.nih.gov/11237119/).
13. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International journal of epidemiology*. 2013; 42(1):111–27. doi: [10.1093/ije/dys064](https://doi.org/10.1093/ije/dys064) PMID: [22507743](https://pubmed.ncbi.nlm.nih.gov/22507743/); PubMed Central PMCID: PMC3600618.
14. Moayyeri A, Hammond CJ, Hart DJ, Spector TD. The UK Adult Twin Registry (TwinsUK Resource). *Twin research and human genetics: the official journal of the International Society for Twin Studies*. 2013; 16(1):144–9. doi: [10.1017/thg.2012.89](https://doi.org/10.1017/thg.2012.89) PMID: [23088889](https://pubmed.ncbi.nlm.nih.gov/23088889/); PubMed Central PMCID: PMC3927054.
15. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010; 26(5):589–95. doi: [10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698) PMID: [20080505](https://pubmed.ncbi.nlm.nih.gov/20080505/); PubMed Central PMCID: PMC2828108.
16. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*. 2007; 81(5):1084–97. doi: [10.1086/521987](https://doi.org/10.1086/521987) PMID: [17924348](https://pubmed.ncbi.nlm.nih.gov/17924348/); PubMed Central PMCID: PMC2265661.
17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38(8):904–9. doi: [10.1038/ng1847](https://doi.org/10.1038/ng1847) PMID: [16862161](https://pubmed.ncbi.nlm.nih.gov/16862161/).
18. International HapMap Consortium, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467(7311):52–8. doi: [10.1038/nature09298](https://doi.org/10.1038/nature09298) PMID: [20811451](https://pubmed.ncbi.nlm.nih.gov/20811451/); PubMed Central PMCID: PMC3173859.
19. Myers GL, Kimberly MM, Waymack PP, Smith SJ, Cooper GR, Sampson EJ. A reference method laboratory network for cholesterol: a model for standardization and improvement of clinical laboratory measurements. *Clinical chemistry*. 2000; 46(11):1762–72. PMID: [11067811](https://pubmed.ncbi.nlm.nih.gov/11067811/).
20. Warnick GR. Laboratory measurement of lipid and lipoprotein risk factors. *Scandinavian journal of clinical and laboratory investigation Supplementum*. 1990; 198:9–19. PMID: [2189213](https://pubmed.ncbi.nlm.nih.gov/2189213/).
21. Zhai G, Wang-Sattler R, Hart DJ, Arden NK, Hakim AJ, Illig T, et al. Serum branched-chain amino acid to histidine ratio: a novel metabolomic biomarker of knee osteoarthritis. *Annals of the rheumatic diseases*. 2010; 69(6):1227–31. doi: [10.1136/ard.2009.120857](https://doi.org/10.1136/ard.2009.120857) PMID: [20388742](https://pubmed.ncbi.nlm.nih.gov/20388742/).
22. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015; 4:7. doi: [10.1186/s13742-015-0047-8](https://doi.org/10.1186/s13742-015-0047-8) PMID: [25722852](https://pubmed.ncbi.nlm.nih.gov/25722852/); PubMed Central PMCID: PMC4342193.
23. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*. 2011; 39(17):e118. doi: [10.1093/nar/gkr407](https://doi.org/10.1093/nar/gkr407) PMID: [21727090](https://pubmed.ncbi.nlm.nih.gov/21727090/); PubMed Central PMCID: PMC3177186.
24. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*. 2013; 41(Database issue):D808–15. doi: [10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094) PMID: [23203871](https://pubmed.ncbi.nlm.nih.gov/23203871/); PubMed Central PMCID: PMC3531103.

25. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics*. 2012; 91(2):224–37. doi: [10.1016/j.ajhg.2012.06.007](https://doi.org/10.1016/j.ajhg.2012.06.007) PMID: [22863193](https://pubmed.ncbi.nlm.nih.gov/22863193/); PubMed Central PMCID: PMC3415556.
26. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS genetics*. 2015; 11(4):e1005165. doi: [10.1371/journal.pgen.1005165](https://doi.org/10.1371/journal.pgen.1005165) PMID: [25906071](https://pubmed.ncbi.nlm.nih.gov/25906071/); PubMed Central PMCID: PMC4407972.
27. Frohne IH, R.J. Sample Quantiles. R Project. ISBN 3-900051-07-0. 2009.
28. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. 2014.
29. R Core Development Team. R Core Team (2013) R: A Language and Environment for Statistical Computing. 2013.
30. Peterson TA, Park D, Kann MG. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC genomics*. 2013; 14 Suppl 3:S5. doi: [10.1186/1471-2164-14-S3-S5](https://doi.org/10.1186/1471-2164-14-S3-S5) PMID: [23819456](https://pubmed.ncbi.nlm.nih.gov/23819456/); PubMed Central PMCID: PMC3665522.
31. Mata IF, Wedemeyer WJ, Farrer MJ, Taylor JP, Gallo KA. LRRK2 in Parkinson's disease: protein domains and functional insights. *Trends in neurosciences*. 2006; 29(5):286–93. doi: [10.1016/j.tins.2006.03.006](https://doi.org/10.1016/j.tins.2006.03.006) PMID: [16616379](https://pubmed.ncbi.nlm.nih.gov/16616379/).
32. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. *Trends in genetics: TIG*. 2013; 29(5):318–27. doi: [10.1016/j.tig.2013.01.004](https://doi.org/10.1016/j.tig.2013.01.004) PMID: [23415593](https://pubmed.ncbi.nlm.nih.gov/23415593/).
33. Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. *Cancer research*. 2008; 68(6):1675–82. doi: [10.1158/0008-5472.CAN-07-5283](https://doi.org/10.1158/0008-5472.CAN-07-5283) PMID: [18339846](https://pubmed.ncbi.nlm.nih.gov/18339846/)
34. Gough J. The SUPERFAMILY database in structural genomics. *Acta crystallographica Section D, Biological crystallography*. 2002; 58(Pt 11):1897–900. PMID: [12393919](https://pubmed.ncbi.nlm.nih.gov/12393919/).
35. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, et al. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic acids research*. 2005; 33(Database issue):D284–8. doi: [10.1093/nar/gki078](https://doi.org/10.1093/nar/gki078) PMID: [15608197](https://pubmed.ncbi.nlm.nih.gov/15608197/); PubMed Central PMCID: PMC540032.
36. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*. 2014; 42(Database issue):D358–63. doi: [10.1093/nar/gkt1115](https://doi.org/10.1093/nar/gkt1115) PMID: [24234451](https://pubmed.ncbi.nlm.nih.gov/24234451/); PubMed Central PMCID: PMC3965093.
37. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An Integrative Approach to Predicting the Functional Effects of Non-Coding and Coding Sequence Variation. *Bioinformatics*. 2015. doi: [10.1093/bioinformatics/btv009](https://doi.org/10.1093/bioinformatics/btv009) PMID: [25583119](https://pubmed.ncbi.nlm.nih.gov/25583119/).
38. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2015; 31(5):761–3. doi: [10.1093/bioinformatics/btu703](https://doi.org/10.1093/bioinformatics/btu703) PMID: [25338716](https://pubmed.ncbi.nlm.nih.gov/25338716/); PubMed Central PMCID: PMC394341060.
39. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. *Nat Genet*. 2014; 46(2):200–4. doi: [10.1038/ng.2852](https://doi.org/10.1038/ng.2852) PMID: [24336170](https://pubmed.ncbi.nlm.nih.gov/24336170/); PubMed Central PMCID: PMC3939031.
40. Surakka I, Horikoshi M, Magi R, Sarin AP, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. *Nat Genet*. 2015; 47(6):589–97. doi: [10.1038/ng.3300](https://doi.org/10.1038/ng.3300) PMID: [25961943](https://pubmed.ncbi.nlm.nih.gov/25961943/).
41. Timpson NJ, Walter K, Min JL, Tachmazidou I, Malerba G, Shin SY, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature communications*. 2014; 5:4871. doi: [10.1038/ncomms5871](https://doi.org/10.1038/ncomms5871) PMID: [25225788](https://pubmed.ncbi.nlm.nih.gov/25225788/); PubMed Central PMCID: PMC4167609.
42. Demirkan A, Amin N, Isaacs A, Jarvelin MR, Whitfield JB, Wichmann HE, et al. Genetic architecture of circulating lipid levels. *European journal of human genetics: EJHG*. 2011; 19(7):813–9. doi: [10.1038/ejhg.2011.21](https://doi.org/10.1038/ejhg.2011.21) PMID: [21448234](https://pubmed.ncbi.nlm.nih.gov/21448234/); PubMed Central PMCID: PMC3137496.
43. Rader DJ, Cohen J, Hobbs HH. Monogenic hypercholesterolemia: new insights in pathogenesis and treatment. *The Journal of clinical investigation*. 2003; 111(12):1795–803. doi: [10.1172/JCI18925](https://doi.org/10.1172/JCI18925) PMID: [12813012](https://pubmed.ncbi.nlm.nih.gov/12813012/); PubMed Central PMCID: PMC161432.
44. Soutar AK, Naoumova RP. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nat Clin Pract Cardiovasc Med*. 2007; 4(4):214–25. doi: [10.1038/npcardio0836](https://doi.org/10.1038/npcardio0836) PMID: [17380167](https://pubmed.ncbi.nlm.nih.gov/17380167/).
45. Feng S, Liu D, Zhan X, Wing MK, Abecasis GR. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics*. 2014; 30(19):2828–9. doi: [10.1093/bioinformatics/btu367](https://doi.org/10.1093/bioinformatics/btu367) PMID: [24894501](https://pubmed.ncbi.nlm.nih.gov/24894501/); PubMed Central PMCID: PMC394173011.