EJHG Open

European Journal of Human Genetics (2017) 25, 123–129
Official journal of The European Society of Human Genetics
www.nature.com/ejhg

ARTICLE

A pathway-centric approach to rare variant association analysis

Tom G Richardson[1], Nicholas J Timpson[1], Colin Campbell[2] and Tom R Gaunt*,[1]

Current endeavours in rare variant analysis are typically underpowered when investigating association signals from individual genes. We undertook an approach to rare variant analysis which utilises biological pathway information to analyse functionally relevant genes together. Conventional filtering approaches for rare variant analysis are based on variant consequence and are therefore confined to coding regions of the genome. Therefore, we undertook a novel approach to this process by obtaining functional annotations from the Combined Annotation Dependent Depletion (CADD) tool, which allowed potentially deleterious variants from intronic regions of genes to be incorporated into analyses. This work was undertaken using whole-genome sequencing data from the UK10K project. Rare variants from the KEGG pathway for arginine and proline metabolism were collectively associated with systolic blood pressure ($P=3.32\text{x}10^{-5}$) based on analyses using the optimal sequence kernel association test. Variants along this pathway also showed evidence of replication using imputed data from the Avon Longitudinal Study of Parents and Children cohort ($P=0.02$). Subsequent analyses found that the strength of evidence diminished when analysing genes in this pathway individually, suggesting that they would have been overlooked in a conventional gene-based analysis. Future studies that adopt similar approaches to investigate polygenic effects should yield value in better understanding the genetic architecture of complex disease.

European Journal of Human Genetics (2017) **25,** 123–129; doi:10.1038/ejhg.2016.113; published online 31 August 2016

INTRODUCTION

Genome-wide association studies (GWAS) have been successful in identifying novel susceptibility loci across a range of complex diseases. However, the underlying mechanisms by which associated variants influence disease or quantitative phenotypes are often undefined and the variance that they explain is almost always small. Next-generation sequencing (NGS) has allowed rarer genetic variants (that is, minor allele frequency (MAF) < 1%), which may potentially contribute substantively to phenotypic variance,[1–4] to be genotyped and imputed in larger samples than before. The analysis of these variants, potentially with larger effect sizes,[1] may prove vital in understanding this missing heritability and uncovering the genetic architecture of complex disease.

Study designs over the last decade have largely concerned the genetic contribution of common variants and thus lack power when applied to rare variants. In order to improve power when analysing these variants using NGS data, methodology based on their aggregation has been developed. Some of these techniques are based on dispersion (c-alpha,[5] SKAT[6]) and are more robust when analysing rare variants that have conflicting directions of effect for a phenotype of interest (protective or deleterious). Despite this, rare variant analyses remain underpowered, owing to inadequate samples sizes and the manner in which the genome is partitioned. One way in which statistical power can potentially be increased is by analysing combinations of variants which have a similar functional impact from within genes that lie along the same molecular pathway.

Such an alternative approach is attractive for several reasons, including a reduction on the burden for multiple testing and identification of association signals contributed to by multiple genes, which would have otherwise been overlooked using single gene approaches. Expanding the region of interest to collapse variants together across multiple genes also has its challenges, such as choice of pathway definitions. Clearly, aggregating variants together from genes which are not functionally relevant will hinder analyses rather than enrich them. Several studies have already attempted to analyse NGS data with pathway or gene network-based approaches, although generally their findings have been underwhelming.[7–9] However, the prospect for developing an approach that successfully identifies polygenic effects from rare variants is potentially very rewarding.

This study aimed to undertake a novel approach to rare variant analysis by focusing on their aggregation across predicted biological pathways, rather than individual loci. Hypothetically, analysing variants from within collective regions may provide stronger evidence of association with phenotypic outcome than single gene approaches. Furthermore, we wanted findings to allow biologically meaningful inferences to be made with regards to the mechanisms of complex disease, which would not be possible when analysing individual genes in isolation. Follow-up analyses of signals detected at a pathway level were performed in order to isolate the evidence of association back to individual genes and variants. We undertook this analysis using whole-genome sequencing (WGS) from participants involved in the UK10K project and cardiovascular phenotypes as measured in the Avon Longitudinal Study of Parents and Children (ALSPAC)[10] and TwinsUK[11] cohorts.

[1]MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, UK; [2]Intelligent Systems Laboratory, University of Bristol, Bristol, UK
*Correspondence: Dr TR Gaunt, MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Bristol BS8 2BN, UK. Tel: +44 117 3310132; Fax: +44 117 3310123; E-mail: Tom.Gaunt@bristol.ac.uk

Received 28 September 2015; revised 18 May 2016; accepted 19 July 2016; published online 31 August 2016

MATERIALS AND 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 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[10,12] (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.[11] Initially, only middle-aged women were recruited and as a result 83% of the registry is female. The registry currently contains 51% monozygotic and 49% dizygotic 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 NRES Committee, Westminster, London.

Sequencing data

DNA Samples from 4030 UK10K study participants (2040 offspring from the ALSPAC cohort, 1990 from the TwinsUK cohort) were subjected to low coverage (6–8 × average read depth) WGS. Sequencing was performed at both the WTSI and the Beijing Genomics Institute (BGI). DNA (1–3 μg) 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).[13] Of the 4030 participants, 3910 samples (1976 ALSPAC and 1934 TwinsUK) went through the variant calling procedure. Low-quality samples were identified by comparing the samples to their GWAS genotypes using ~20 000 sites on chromosome 20. A total of 112 samples (48 ALSPAC and 64 TwinsUK) were removed, leaving 3798 samples (1928 ALSPAC and 1870 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[14] 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 owing 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[15] to exclude participants of non-European ancestry after merging our data with a pruned 11 HapMap3 population data set.[16] Forty-four subjects (12 TwinsUK and 32 ALSPAC) did not cluster to the European cluster and were removed. The final sample size for association analyses comprised of 3621 individuals (1754 TwinsUK and 1867 ALSPAC).

UK10K phenotypes

ALSPAC. Body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP) and non-fasting blood samples were obtained from the age 9 clinic (mean age: 9.9, range: 8.9–11.5). The blood samples were used to determine lipids including total cholesterol (TC), triglycerides (TG) and high-density lipoproteins (HDL) using a modification of the standard Lipid Research Clinic Protocol.[17] Low-density lipoproteins were subsequently calculated using the Friedwald equation:[18]

TwinsUK. BMI, SBP, DBP and blood samples taken after at least 6 h of fasting were obtained from study participants (mean age: 56, range: 17–85). 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.

Further details on phenotype measurements in the UK10K consortium are provided in the Supplementary Material (Supplementary Information 1).

ALSPAC imputed data for replication analyses

Genome-wide association (GWA) data was also available for 6557 ALSPAC samples who were not enroled in the UK10K project. This data was genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform (Illumina Inc, San Diego, CA, USA) by the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK) and the Laboratory Corporation of America (LCA, Burlington, NC, USA). All individuals passed QC based on incorrect sex assignment; abnormal heterozygosity (<0.320 or >0.345 for WTSI data; <0.310 or >0.330 for LCA data); high missingness ($>3\%$); cryptic relatedness ($>10\%$ identity by descent) and non-European ancestry (detected by multidimensional scaling analysis).

Variants discovered through WGS in the UK10K project were used for the development and use of a reference panel for this imputation within the ALSPAC and TwinsUK GWA data sets. These were combined with known variants taken from the 1000 Genomes reference panel. Novel functionality was developed in IMPUTE2[19,20] to use each reference panel to impute missing variants in their counterparts before ultimately combining them together. This meant that all loci involved in the discovery analysis were imputed in this sample. Further detail on this imputation process can be found in the publication by Timpson *et al.*[20] All rare variants (MAF $\leq$ 1%) were filtered to have Hardy–Weinberg equilibrium $P > 5 \times 10^{-7}$. As accurate imputation of rare variants can be challenging, only variants with an imputation quality score of 0.4 or higher were used in this analysis. Phenotype measurements were collected as before for lipid traits, although for BMI, SBP and DBP we used measurements obtained from the focus at age 7 clinic (mean age: 7.5, range: 7.1–8.8) in order to maximise sample size.

Please note that the study website contains details of all the data that are available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary). Ethical approval was obtained from the ALSPAC Law and Ethics Committee and the Southmead, Frenchay, UBHT and Weston Research Ethics Committees. Written informed consent was obtained from parents for all measurements made.

Statistical analysis

BioCarta, KEGG and Reactome pathways analyses with cardiovascular traits. Pathway definitions were obtained from three pathway databases (BioCarta,[21] KEGG[22] and Reactome[23]). This was achieved by downloading gene sets from the Broad institute's molecular signatures database (MSigDB, found at http://www.broadinstitute.org/gsea/downloads.jsp) which is used by the GSEA algorithm.[24] All rare genetic variants (MAF $\leq$ 1%) that survived QC were aggregated together across pathways according to UCSC reference genome hg19 definitions. We then filtered these sets to only include nonsynonymous variants according to the Ensembl Variant Effect Predictor (VEP).[25] This process was repeated for a separate analysis except using variants resulting in a loss-of-function (that is, stop gains/losses, splice sites or frameshift indels according to VEP annotations). In a final set of analyses, we used the Combined Annotation Dependent Depletion (CADD)[26] resource to filter variants based on molecular functionality and pathogenicity. Based on simulations using network analyses surrounding candidate loci (Supplementary Information 2 and Supplementary Figure 1), we hypothesised that a strict threshold would be necessary for this analysis and therefore only included the top 1% most deleterious variants in the UK10K WGS data as predicted by CADD (that is, all variants with a CADD C-Score $\geq$ 20).

Rare variants (MAF $\leq$ 1%) from these regions were then analysed with each cardiovascular trait using the optimal sequence kernel association test (SKAT-O).[27] Adjustment for the first 10 ancestry principal components was undertaken for all analyses as it has been reported that only adjusting for a small number of top principal components when analysing rare variants can lead to inflated type 1 error rates.[28] Screeplots suggested that this should be sufficient to account for any potential population stratification in the planned analyses (Supplementary Figure 2).

Multiple comparisons and permutation testing. Results that survived the conservative multiple testing threshold (0.05/# of pathways analysed) were further evaluated by performing 100 000 permutation tests. For each test, the same set of genotype values was reanalysed as before except after randomly shuffling phenotype values among individuals. *P*-values were evaluated as robust to permutation if no more than 10 permuted *P*-values were smaller than the *P*-values observed when phenotypes were not shuffled (that is, the observed association).

Follow-up analyses to isolate association signals. We followed up pathways that were robust to permutation testing by analysing all genes individually across the genome using the appropriate variant filtering technique and corresponding phenotype. Results were plotted using Manhattan plots, highlighting genes located along the pathway of interest.

We also examined protein–protein interactions to elucidate subnetworks of genes within any pathways with strong evidence of association in the previous analysis. These subnetworks were determined due to the most connected genes in the pathway based on STRINGdb experimental data. Circos plots[29] were used to visualise how these subnetworks interacted across the genome. Subnetworks of genes were analysed in a similar manner to before in order to isolate association signals across pathways.

Replication. Using the imputed ALSPAC data set, we attempted to replicate evidence of association for pathways that survived the multiple testing threshold and were robust to permutation testing. All SNVs in pathways found to be associated in the UK10K data (with imputation quality ≥ 0.4) were collapsed together as before and SKAT-O analyses were repeated using the respective trait. SNVs with an imputation quality score less than this were not included in this analysis. All SNVs in this analysis were also collapsed according to individual gene coordinates to see if there was any overlap with key genes identified in the discovery analysis. As before, analyses were adjusted for the top 10 principal components. Single variant analyses were also undertaken using PLINKv1.9[30] to verify whether there were any key SNVs in both sets of analyses. All other statistical analyses in this study were undertaken using R statistical software.[31]

RESULTS

Figure 1 shows a flowchart of the planned analysis for this study. In total, 3621 individuals with WGS data from the UK10K cohort were available for analysis. After merging individuals with each cardiovascular trait that we planned to analyse, final sample sizes ranged between 3538 and 3191 (3538 for BMI, 3309 for SBP and DBP, 3210 for HDL, 3191 for LDL, 3206 for TC and 3202 for TG).

KEGG, Reactome and BioCarta pathways analyses with cardiovascular traits

In total, 1095 pathway definitions were obtained from the three pathway databases. We observed no strong evidence of association for any pathway results after filtering to only include nonsynonymous variants based on the multiple testing correction threshold ($0.05/1095 = P < 4.57 \times 10^{-5}$). Results can be located in Supplementary Table 1. There were four results that survived the multiple testing correction threshold in the loss-of-function analysis with TG ($0.05/897 = P < 5.57 \times 10^{-5}$). However, all these signals were driven by a single rare variant located within the *APOC3* gene region (rs138326449), which has been previously identified in the UK10K data set.[20] These results can be located in Supplementary Table 2.

The following results for our study now solely concern the analysis based on filtering variants based on their CADD score. One pathway result survived the conservative correction for multiple testing ($P < 4.65 \times 10^{-5}$, that is, 0.05/1.075 based on the number of pathways with two or more rare variants after filtering by a CADD C-Score threshold of ≥ 20), which was associated with SBP. This was the

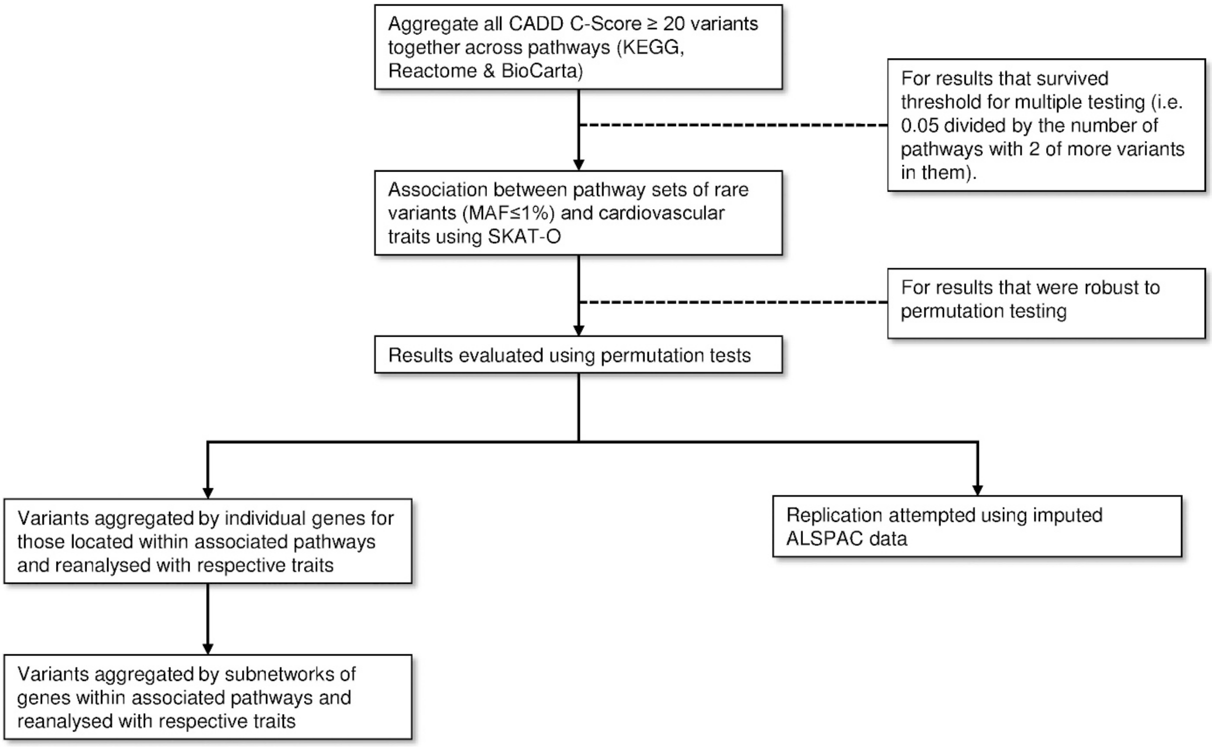
Aggregate all CADD C-Score ≥ 20 variants together across pathways (KEGG, Reactome & BioCarta)
For results that survived threshold for multiple testing (i.e. 0.05 divided by the number of pathways with 2 of more variants in them).
Association between pathway sets of rare variants (MAF≤1%) and cardiovascular traits using SKAT-O
For results that were robust to permutation testing
Results evaluated using permutation tests
Variants aggregated by individual genes for those located within associated pathways and reanalysed with respective traits
Replication attempted using imputed ALSPAC data
Variants aggregated by subnetworks of genes within associated pathways and reanalysed with respective traits

Figure 1 A flowchart illustrating the pipeline for analysis in this study. CADD, combined annotation dependent depletion tool; SKAT-O, optimal sequence kernel association tool; ALSPAC, avon longitudinal study of parents and children.

Table 1 Systolic blood pressure results using a minor allele frequency cutoff of 1%

Pathway	*Variants*[a]	*UK10K* P-*value*	*TwinsUK* P-*value*	*ALSPAC* P-*value*
KEGG: arginine and proline metabolism	115	3.32×10^{-5}	7.42×10^{-4}	0.04
Reactome: amino acid synthesis and interconversion transamination	26	9.69×10^{-4}	0.04	0.06
KEGG: antigen processing and presentation	18	5.78×10^{-3}	0.27	0.02
Reactome: synthesis secretion and deacylation of ghrelin	13	6.12×10^{-3}	0.04	0.13
KEGG: tight junction	412	8.48×10^{-3}	0.29	0.36
KEGG: Jak Stat signalling pathway	90	0.01	0.34	0.04
BioCarta: cytokine pathway	5	0.01	1.91×10^{-3}	0.48
Reactome: regulation of IFNG signalling	20	0.01	0.38	0.02
Reactome: assembly of the pre replicative complex	97	0.02	0.12	0.12
BioCarta: RAB pathway	9	0.02	0.15	0.05

[a] Variants = number of variants, UK10K *P*-value = SKAT-O *P*-value for entire UK10K sample, Twins *P*-value, SKAT-O *P*-value for only TwinsUK individuals, ALSPAC *P*-value = SKAT-O *P*-value for only ALSPAC individuals.

KEGG arginine and proline Metabolism pathway ($P=3.32\times10^{-5}$), referred to hereafter as the arginine and proline pathway. The top results from this analysis can be located in Table 1. The strongest evidence of association in the BMI analysis was with variants along the *IGF1* pathway according to BioCarta, although this did not survive the threshold for multiple testing ($P=2.88\times10^{-4}$). The top results for this analysis and all other traits can be found in Supplementary Tables 3.

Permutation testing

No permuted *P*-value survived the multiple testing threshold based on the 100 000 iterations undertaken. Furthermore, only three permuted *P*-values were lower than the observed association between the arginine and proline pathway and SBP (Supplementary Table 4). These results supported evidence that this effect was not due to chance.

Follow-up analyses to isolate association signals

We collapsed all rare variants (MAF ≤ 1%) across the genome with a CADD C-Score threshold of ≥ 20 within individual gene regions according to coordinates based on UCSC reference genome hg19 definitions. Each gene-based set of variants was analysed using SKAT-O with SBP. Figure 2 shows a Manhattan plot for the results of this analysis. Genes located along the arginine and proline pathway are highlighted. Notably, none of these *P*-values survived the threshold for multiple comparisons ($P<6.05\times10^{-6}$ (0.05/8269, based on the 8269 gene regions with two or more variants analysed)). This shows that a conventional gene-based approach would not have been able to detect the effect across rare variants in the arginine and proline pathway. Supplementary Table 5 show the results for this individual gene analysis for genes located on this pathway.

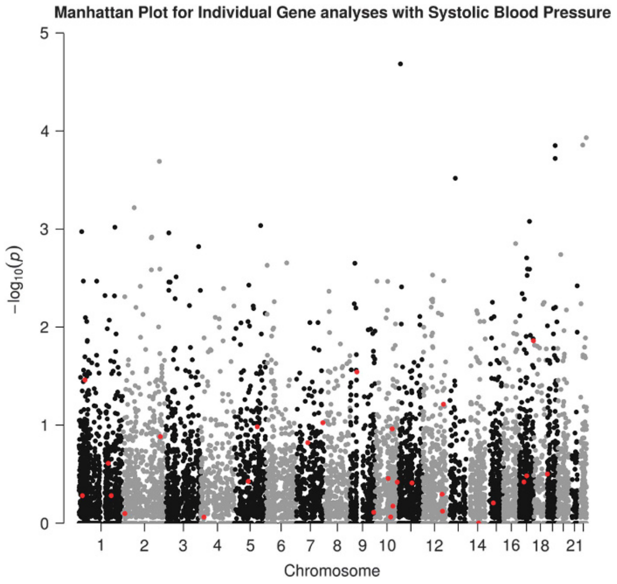

Figure 2 A Manhattan plot illustrating gene-based collapsing of rare variants (MAF ≤ 1%) genome-wide using SKAT-O with systolic blood pressure. Red annotated points represent genes which reside along the arginine and proline pathway.

We also examined protein–protein interactions according to Stringdb[32] experimental data to identify subnetworks of genes within the arginine and proline pathway. There were three genes within this pathway (*CPS1, PRODH2* and *GLUD1*), which had more interactions than any others and therefore subnetworks were based around all genes directly connected to them. Figure 3 displays a Circos plot,[29] which illustrates interactions between subnetworks within this pathway. Variants within these subnetworks were analysed as before using SKAT-O with SBP. Two subnetworks (the *CPS1* and *GLUD1* subnetworks) provided stronger evidence of association over any individual gene analysis ($P=1.58\times10^{-4}$ and 1.30×10^{-3}, respectively), however only when all genes in the network were analysed together did evidence become strong enough to survive the threshold for multiple testing in the initial analysis ($P=4.65\times10^{-5}$). Table 2 shows the full results of this analysis.

Replication using ALSPAC imputed data

We undertook a replication analysis using imputed data from the ALSPAC data set using individuals who were not enroled in the UK10K project (that is, data from a completely separate population of individuals than those already analysed). We replicated evidence of association in the imputed data set between variants along the arginine and proline pathway and SBP ($P=0.02$ from 71 variants, $N=4380$). Evidence of association appeared to be consistent from both the discovery and replication analyses for variants within the *NOS1* gene ($P=0.06$ and $P=0.05$, respectively), although there was little concordance between key individual variant results between the discovery and replication sets of analyses (Supplementary Tables 6–8).

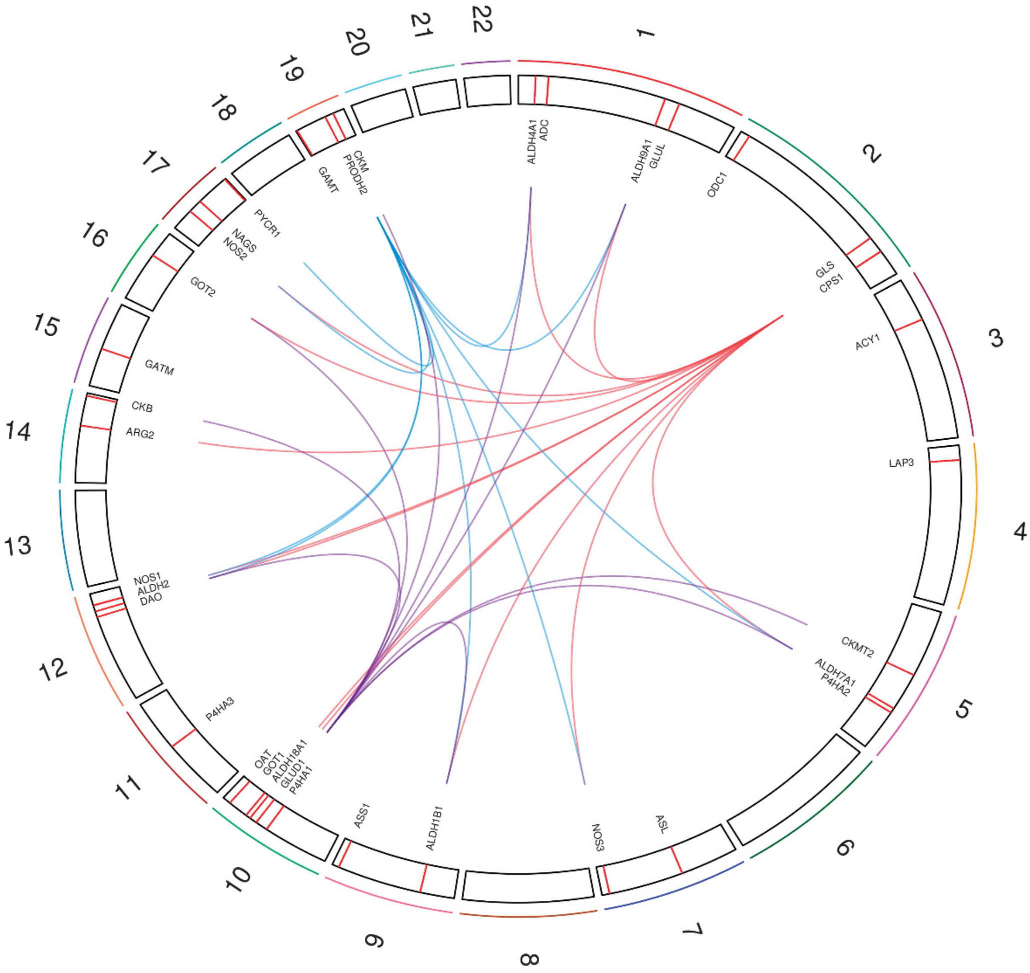

Figure 3 Circos plot representing the genomic location of Genes in the Arginine and Proline pathway and how their protein products interact according to Stringdb. Stringdb suggested that the protein products of three genes along the Arginine and Proline pathway had the most interactions based on experimental evidence. The following three subnetworks are based on the interactions for each of these genes (*CPS1, PRODH2* and *GLUD1*): *CPS1, NOS1, NOS2, NOS3, ALDH9A1, ARG2, ALDH18A1, ALDH1B1, ALDH2, ALDH7A1, ALDH4A1, GOT1* and *GOT2. PRODH2, NOS1, NOS2, NOS3, PYCR1, ALDH9A1, ALDH4A1, ALDH2, ALDH7A1* and *ALDH1B1. GLUD1, CKM, CKMT2, CKB, ALDH1B1, ALDH7A1, ALDH2, ALDH9A1, ALDH4A1* and *GOT2.* Red, blue and purple links are used to differentiate the three subnetworks.

Table 2 Subnetwork analysis for the KEGG arginine and proline metabolism pathway with systolic blood pressure

	UK10K		*TwinsUK*		*ALSPAC*	
Genes in Subnetwork	*Variants*[a]	*P-value*	*Variants*[a]	*P-value*	*Variants*[a]	*P-value*
CPS1, NOS1, NOS2, NOS3, ALDH9A1, ARG2, ALDH18A1, ALDH1B1, ALDH2, ALDH7A1, ALDH4A1, GOT1 and *GOT2*	57	1.58×10^{-4}	38	4.00×10^{-3}	40	0.03
PRODH2, NOS1, NOS2, NOS3, PYCR1, ALDH9A1, ALDH4A1, ALDH2, ALDH7A1 and *ALDH1B1*	25	0.02	18	0.16	20	0.15
GLUD1, CKM, CKMT2, CKB, ALDH1B1, ALDH7A1, ALDH2, ALDH9A1, ALDH4A1 and *GOT2*	34	1.30×10^{-3}	25	0.02	26	0.06

[a] Variants = number of variants analysed, UK10K = all individuals within UK10K project, TwinsUK = only TwinsUK partition of the UK10K project, ALSPAC = only ALSPAC partition of the UK10K project, *P*-value = *P*-values according to SKAT-O test.

DISCUSSION

We present an approach to rare variant analysis that expands the region of interest from individual genes to established biological pathways. This approach uses a variant filtering method that incorporates functional annotations from the CADD tool, which allows both intronic and exonic variants to be included into analyses. This is in contrast to conventional filtering methods in rare variant analysis, which are based on variant consequence and are therefore confined to coding regions. Implementing this approach has identified evidence to suggest a set of functionally relevant rare variants collectively influence SBP. Evaluations of this pathway identified subsets of genes that were predominantly responsible for the observed signal, although only when analysed as an entire pathway did statistical evidence become robust enough to survive the correction for multiple comparisons.

The observed association between genes involved in the metabolism of arginine and proline and SBP ($P = 3.32\times10^{-5}$ from 115 variants,

$N = 3309$) in the UK10K data set appeared to replicate in the imputed ALSPAC data set ($P = 0.02$ from 71 variants, $N = 4380$). There is evidence to suggest that both these amino acids can influence blood pressure, although arginine in particular has gained interest with several studies reporting a blood pressure lowering effect due to arginine dietary supplements.[33–35] An important function of arginine is a necessary precursor for the synthesis of nitric oxide, where the impairment of nitric oxide bioavailability has been shown to be a risk factor for hypertension.[36] *NOS1* was observed to be one of the key genes in both the discovery and replication analyses within this study. This gene belongs to the family of nitric oxide synthases and is involved in the synthesis of nitric oxide from arginine, making it an interesting candidate to explore the mechanisms within this pathway and its effect on blood pressure.

We observed a lack of concordance between key individual variants in the discovery and replication analyses, although this can most likely be explained by the typically observed lack of power when conducting single variant analyses on rare variants.[37,38] Replication in another cohort was also contemplated but not undertaken owing to the challenge of identifying a cohort where a sufficient proportion of variants within the discovery analysis were genotyped. Moreover, variants at this frequency are more likely to be population specific in comparison with replication for GWAS hits, which makes replication of rare variant signals challenging, particularly in a trans-ethnic context.[39] An attractive alternative therefore to the manner in which rare variants have been analysed in this study is to meta-analyse results using software such as RAREMETAL.[40]

As larger sample sizes and lower coverage for sequence data become available this approach should also provide a powerful approach to uncover the role of extremely rare genetic variants (MAF $\leq 0.5\%$). For future applications of this approach we advise implementing a strict variant filtering phase before analysis in order to remove as much statistical noise as possible. In our initial analysis when filtering variants by conventional methods based on variant consequence we did not observe any novel evidence of association in the UK10K data set. However, when incorporating annotations from CADD we were able to identify a previously undetected signal. Although other studies have incorporated non-coding annotations into their analyses to great effect,[41] this study presents an approach for utilising them in the context of rare variant analysis within a pathway-centric framework. This is a noteworthy point as when performing a conventional gene-based analysis evidence of association for this novel signal diminished.

The limitation of this approach is that potential causal variants may be filtered out when based on CADD predictions. In this study we have chosen pathway definitions using the KEGG, Reactome and BioCarta resources, although future studies may benefit from incorporating gene sets based on alternative definitions. We have also showcased SKAT-O as an adaptive method suited to this approach. Along with its capability to identify phenotype associated regions of variation, which can have conflicting directions of effect, the manner in which SKAT-O adjusts for LD, using the Davies method,[42] also makes it an attractive approach when analysing multiple genes in a pathway which can potentially have close proximity. As reported by other studies, analysing pathway sets of variants using SKAT-O, or with any other collapsing method, remains relatively underpowered.[7,8] Power calculations from the replication analysis in this study also support this, as evaluations found observed power to diminish as the sample size and percentage of variants in the sample were reduced (Supplementary Information 3).

Success from rare variant association analyses has been limited thus far.[43,44] In recent years, studies have attempted to improve statistical power in their approaches by increasing their sample sizes,[45] adapting current methodological techniques[46] or conducting analyses on remote or isolated populations where the allele frequency of rarer variants may be heightened.[47] By expanding our definition of a biologically functional unit from individual genes, we have provided an alternative method to leverage statistical power. Furthermore, subsequent analysis concerning pathway enrichment (for example, GSEA[24]) would struggle to detect an over representation of genes along certain pathways. Our approach allows pathological mechanistic insights to be made which methods involving individual gene regions would not. This is also an advantage over individual variant methods, such as GWAS, where the majority of robustly associated findings lie in non-coding regions, leaving questions regarding the variants' functional impact.[48] As larger samples of low coverage sequence data become available, future studies should benefit from using this approach to uncover the genetic architecture of complex traits that contribute to disease.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

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. 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. GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23 and Me. 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.

1 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; **40**: 695–701.

2 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.

3 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 2004; **305**: 869–872.

4 Cohen JC, Pertsemlidis A, Fahmi S *et al*: Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA* 2006; **103**: 1810–1815.

5 Neale BM, Rivas MA, Voight BF *et al*: Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; **7**: e1001322.

6 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. *Am J Hum Genet* 2011; **89**: 82–93.

7 Aslibekyan S, Almeida M, Tintle N: Pathway analysis approaches for rare and common variants: insights from Genetic Analysis Workshop 18. *Genet Epidemiol* 2014; **38**: S86–S91.

8 Zhao J, Zhu Y, Boerwinkle E, Xiong M: Pathway analysis with next-generation sequencing data. *Eur J Hum Genet* 2015; **23**: 507–515.

9 Hu P, Paterson AD: Dynamic pathway analysis of genes associated with blood pressure using whole genome sequence data. *BMC Proc* 2014; **8**: S106.

10 Boyd A, Golding J, Macleod J *et al*: Cohort Profile: the 'children of the 90 s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* 2013; **42**: 111–127.

11 Moayyeri A, Hammond CJ, Hart DJ, Spector TD: The UK adult twin registry (TwinsUK Resource). *Twin Res Hum Genet* 2013; **16**: 144–149.

12 Golding J, Pembrey M, Jones R, Team AS: ALSPAC–the Avon longitudinal study of parents and children. I. study methodology. *Paediatr Perinat Epidemiol* 2001; **15**: 74–87.
13 Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010; **26**: 589–595.
14 Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084–1097.
15 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**: 904–909.
16 International HapMap Consortium, Altshuler DM, Gibbs RA *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
17 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. *Clin Chem* 2000; **46**: 1762–1772.
18 Warnick GR: Laboratory measurement of lipid and lipoprotein risk factors. *Scand J Clin Lab Invest Suppl* 1990; **198**: 9–19.
19 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
20 Timpson NJ, Walter K, Min JL *et al*: A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat Commun* 2014; **5**: 4871.
21 Nishimura D: BioCarta. *Biotech Softw Internet Rep* 2001; **2**: 117–120.
22 Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
23 Croft D, Mundo AF, Haw R *et al*: The Reactome pathway knowledgebase. *Nucleic Acids Res* 2014; **42**: D472–D477.
24 Subramanian A, Tamayo P, Mootha VK *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**: 15545–15550.
25 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010; **26**: 2069–2070.
26 Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**: 310–315.
27 Lee S, Emond MJ, Bamshad MJ *et al*: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genetics* 2012; **91**: 224–237.
28 Zhang Y, Guan W, Pan W: Adjustment for population stratification via principal components in association analysis of rare variants. *Genetic Epidemiol* 2013; **37**: 99–109.
29 Krzywinski M, Schein J, Birol I *et al*: Circos: an information aesthetic for comparative genomics. *Genome Res* 2009; **19**: 1639–1645.
30 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.
31 Core R Development Team: R Core Team (2013) R: A Language and Environment for Statistical Computing 2013.
32 Franceschini A, Szklarczyk D, Frankild S *et al*: STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2013; **41**: D808–D815.
33 Rytlewski K, Olszanecki R, Korbut R, Zdebski Z: Effects of prolonged oral supplementation with l-arginine on blood pressure and nitric oxide synthesis in preeclampsia. *Eur J Clin Invest* 2005; **35**: 32–37.
34 Palloshi A, Fragasso G, Piatti P *et al*: Effect of oral L-arginine on blood pressure and symptoms and endothelial function in patients with systemic hypertension, positive exercise tests, and normal coronary arteries. *Am J Cardiol* 2004; **93**: 933–935.
35 Cicero AF, Aubin F, Azais-Braesco V, Borghi C: Do the lactotripeptides isoleucine-proline-proline and valine-proline-proline reduce systolic blood pressure in European subjects? A meta-analysis of randomized controlled trials. *Am J Hypertens* 2013; **26**: 442–449.
36 Hermann M, Flammer A, Luscher TF: Nitric oxide in hypertension. *J Clin Hypertens* 2006; **8**: 17–29.
37 Lee S, Abecasis GR, Boehnke M, Lin X: Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 2014; **95**: 5–23.
38 The UK10K Consortium: The UK10K project identifies rare variants in health and disease. *Nature* 2015; **526**: 82–90.
39 Moutsianas L, Morris AP: Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits. *Brief Funct Genomics* 2014; **13**: 362–370.
40 Feng S, Liu D, Zhan X, Wing MK, Abecasis GR: RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* 2014; **30**: 2828–2829.
41 Ionita-Laza I, McCallum K, Xu B, Buxbaum JD: A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016; **48**: 214–220.
42 Davies R: The distribution of a linear combination of chi-square random variables. *J R Stat Soc Ser C Appl Stat* 1980; **29**: 323–333.
43 Ladouceur M, Zheng HF, Greenwood CM, Richards JB: Empirical power of very rare variants for common traits and disease: results from sanger sequencing 1998 individuals. *Eur J Hum Genet* 2013; **21**: 1027–1030.
44 Hunt KA, Mistry V, Bockett NA *et al*: Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 2013; **498**: 232–235.
45 Liu DJ, Peloso GM, Zhan X *et al*: Meta-analysis of gene-level tests for rare variant association. *Nat Genet* 2014; **46**: 200–204.
46 Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN: Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol* 2013; **37**: 409–418.
47 Tachmazidou I, Dedoussis G, Southam L *et al*: A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* 2013; **4**: 2872.
48 Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* 2012; **30**: 1095–1106.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)