PLOS ONE

RESEARCH ARTICLE

Incorporating Non-Coding Annotations into Rare Variant Analysis

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

1 MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, **2** Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom

* Tom.Gaunt@bristol.ac.uk

Citation: Richardson TG, Campbell C, Timpson NJ, Gaunt TR (2016) Incorporating Non-Coding Annotations into Rare Variant Analysis. PLoS ONE 11 (4): e0154181. doi:10.1371/journal.pone.0154181

Editor: Junwen Wang, The University of Hong Kong, HONG KONG

Received: December 15, 2015

Accepted: April 11, 2016

Published: April 29, 2016

Data Availability Statement: Due to the potentially sensitive nature of some of the sample sets used, the restrictions on data usage imposed by some Research Ethics Committees (RECs) and the nature of the existing consents concerning study participation, UK10K data are not openly available. Please see http://www.uk10k.org/assets/24229443.pdf for more details. Data are provided by the data custodians under a "managed data access" mechanism via the European Genome/Phenome Archive (EGA, accession EGAO00000000079), with details about data access on the UK10K website (http://www.uk10k.org/data_access.html).

Abstract

Background

The success of collapsing methods which investigate the combined effect of rare variants on complex traits has so far been limited. The manner in which variants within a gene are selected prior to analysis has a crucial impact on this success, which has resulted in analyses conventionally filtering variants according to their consequence. This study investigates whether an alternative approach to filtering, using annotations from recently developed bioinformatics tools, can aid these types of analyses in comparison to conventional approaches.

Methods & Results

We conducted a candidate gene analysis using the UK10K sequence and lipids data, filtering according to functional annotations using the resource CADD (Combined Annotation-Dependent Depletion) and contrasting results with 'nonsynonymous' and 'loss of function' consequence analyses. Using CADD allowed the inclusion of potentially deleterious intronic variants, which was not possible when filtering by consequence. Overall, different filtering approaches provided similar evidence of association, although filtering according to CADD identified evidence of association between *ANGPTL4* and High Density Lipoproteins (P = 0.02, N = 3,210) which was not observed in the other analyses. We also undertook genome-wide analyses to determine how filtering in this manner compared to conventional approaches for gene regions. Results suggested that filtering by annotations according to CADD, as well as other tools known as FATHMM-MKL and DANN, identified association signals not detected when filtering by variant consequence and vice versa.

Conclusion

Incorporating variant annotations from non-coding bioinformatics tools should prove to be a valuable asset for rare variant analyses in the future. Filtering by variant consequence is only possible in coding regions of the genome, whereas utilising non-coding bioinformatics annotations provides an opportunity to discover unknown causal variants in non-coding

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). TGR is a UK MRC PhD Student.

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

regions as well. This should allow studies to uncover a greater number of causal variants for complex traits and help elucidate their functional role in disease.

Introduction

Genome wide association studies (GWAS) have had a profound influence on the number of disease associated common variants detected across the genome, although recently a greater emphasis has been placed on the impact that rarer variants can have on disease[1, 2]. The advent of next generation sequencing (NGS) has facilitated the development of rare variant association approaches by collapsing variants within the same gene together and analysing their combined effect on phenotypic traits[3]. The aim of current endeavors is to identify deleterious regions of variants with potentially larger effect sizes than those typically identified in GWAS[4], subsequently improving the explained heritability of common diseases. Furthermore, rare variant analyses can help identify genes that are relevant to a particular disease, where evidence gained from single variant approaches may be limited.

Rare variant association studies have had varying degrees of success in recent years[5, 6]. The number of variants with little to no effect (neutral variants) and also variants with contrasting directions of effect within a collapsed region often weaken the statistical power of analyses. As a result, methods based around the analysis of variance-components (e.g. C-Alpha[7], Sequence Kernel Association Test (SKAT)[8]) have been developed to overcome these challenges. Furthermore, filtering variants by their functional consequence (e.g. only analysing variants within a region that are predicted to have a 'nonsynonymous' or 'loss of function' impact) has become common practice to reduce the potential number of neutral variants within a collapsed region or functional unit.

Filtering in this manner can have limitations for several reasons. Firstly, even if a variant is annotated correctly, this does not necessarily mean that it will have a deleterious effect (e.g. if a nonsynonymous variant is located within the transmembrane region of a receptor gene it may not drastically alter the function of the protein). Secondly, it is also possible that other types of variants within the coding region of the gene (i.e. silent mutations) can potentially have a deleterious effect but may get filtered out. Lastly, filtering variants according to their consequence is limited to coding regions, therefore making annotations for filtering non-coding regions (i.e. intronic regions of a gene or intergenic regions of the genome) an attractive commodity, particularly with the current influx of whole genome sequence (WGS) data.

The Combined Annotation-Dependent Depletion (CADD)[9] method objectively integrates a range of different annotation metrics into a single measure (C score). By doing so, CADD aims to provide a more reliable estimate of deleteriousness for all known variants and therefore an overall rank for this metric across the genome. Other bioinformatics tools, such as SIFT [10] and Polyphen-2 [11] have previously been used to filter variants for rare variant analyses [12], although importantly these tools use protein-based metrics and are therefore confined to coding regions. Although CADD has been used to further evaluate the impact of SNPs after identification in association studies, this resource has not yet been utilised to filter variants according to prior knowledge about likely function before undertaking a low frequency or rare variant association analysis. We therefore hypothesised that, in comparison to filtering by variant consequence, using CADD may be more informative in identifying functional variants for a complex trait, whilst keeping the number of neutral variants to a minimum. Moreover, when applying collapsing methods to gene regions, CADD allows analyses to be undertaken in

intronic regions of the gene which can potentially harbour functional variants [13], which would not be possible when filtering by variant consequence or prediction tools confined to coding regions.

Using WGS data from the UK10K project (http://www.uk10k.org), we have undertaken low frequency and rare variant analyses to evaluate whether filtering variants according to CADD scaled C-Scores identifies association signals not detected when filtering according to variant consequence. Our hypothesis was that incorporating non-coding bioinformatics annotations based on predicted variant functionality would be a valuable asset for future studies which conduct these types of approaches.

Results

3,781 whole-genome samples from the UK10K cohort arm [14] were available for analysis (1,927 from ALSPAC, 1,854 from TwinsUK) after variant calling and quality control. After merging individuals with each cardiovascular, final sample sizes ranged between 3,538 and 3,191 (3,538 for Body Mass Index (BMI), 3,309 for Systolic and Diastolic Blood Pressure (SBP & DBP), 3,210 for High Density Lipoproteins (HDL), 3,191 for Low Density Lipoproteins (LDL), 3,206 for Total Cholesterol (TC) and 3,202 for Triglycerides (TG)).

Candidate Gene Analysis

We used CADD[9] to obtain scaled C scores for all 44.9 million possible variants and indels in the UK10K whole genome sequence data. After removing variants which failed QC, we filtered all variants in three ways:

1. Variants responsible for a 'nonsynonymous' substitution according to the Variant Effect Predictor [15] (VEP). VCFtools [16] was subsequently used to condense these regions down to just those variants.
2. Variants responsible for a 'loss of function' according to VEP (i.e. 'stop losses/gains', 'splice sites' or 'frameshift indels'). VCFtools was used again to condense these regions down to just those variants.
3. Variants with a CADD C-Score $\geq$15 (i.e. the 5% most damaging variants predicted across the genome). This is a suggested cutoff by the developers of CADD to identify potentially pathogenic variants as it is the median value for all possible canonical splice site changes and nonsynonymous variants (http://cadd.gs.washington.edu/info).

We collapsed variants together across candidate genes and analysed them using SKAT[8] with their associated traits according to Liu et al[17]. These genes were *ANGPTL4*, *BCAM*, *CBLC*, *CD300LG*, *HNF4A*, *LDLR*, *LIPC*, *LIPG*, *LPL*, *PCSK9* and *PVR*. Analyses were repeated after applying each variant filtering method, as well as applying two different minor allele frequency (MAF) cutoffs of MAF$\leq$5% and MAF$\leq$1%. Importantly, filtering by variant consequence was confined to coding regions, whereas filtering by CADD definitions also allowed the inclusion of potentially deleterious non-coding variants which reside in intronic regions of genes. The results of this analysis, as well as those in subsequent analyses of this study, were not adjusted for multiple comparisons. This was because we were interested in the comparison of filtering approaches in terms of identifying association signals, rather than evaluating whether these signals are real, which is important for analyses regardless of filtering method. Furthermore, the same number of analyses were conducted for each filtering method as this would have otherwise incorporated bias into the study.

Using a MAF cutoff of 5%, we observed evidence of association between *LIPG* and HDL (P = 0.02) as well as between *PVR* and LDL (P = 0.02) after filtering to only include nonsynonymous variants. For the loss of function variant analyses, *CD300LG*, *BCAM* and *ANGPTL4* were associated with HDL, LDL and TG respectively (P = 0.02, P = 0.04 and P = 0.03). Using CADD to filter variants provided evidence of association between three of the previously mentioned genes and traits, as well as between *ANGPTL* and HDL (P = 0.02). The majority of these effects appeared to be driven by rare variants, as evidence of association was observed after applying a MAF cutoff of 1%. The only exception to this was the association between *LIPG* and HDL after filtering to include nonsynonymous variants, which did not provide strong evidence of association using this cutoff (P = 0.07). Tables 1 and 2 show the complete results of our gene-level low frequency variant and rare association analyses, using a cutoff of 5% and 1% MAF respectively.

Genome-wide Analysis of Gene-based Association Signals using CADD

We also identified variants with a CADD C-Score $\geq$15 across the genome and aggregated them together across all gene regions according to UCSC definitions (reference genome hg19). Variants were then collapsed together and analysed with 7 cardiovascular traits (BMI, SBP, DBP, HDL, LDL, TC & TG) using SKAT after applying a MAF cutoff of 5%. We repeated this process in a second set of analyses using a MAF cutoff of 1% to investigate rarer variation. This process was repeated except filtering to include variants which led to a 'nonsynonymous' or 'loss-of-function' consequence (i.e. regardless of CADD C-Score) according to dbSNP annotations (build 137) in two other separate sets of analyses. Only regions which had at least 2 variants remaining after filtering were eligible for analyses, as a single remaining variant would offer no added value when analysed using SKAT compared to using a single variant test.

Table 1. Results of gene-level low frequency variant association tests using various variant filters (MAF $\leq$ 5%).

		Nonsynonymous variants		Loss-of-Function variants		CADD variants (C-Score $\geq$ 15)	
Gene	**Lipid trait**	**nVars**	**P-value**	**nVars**	**P-value**	**nVars**	**P-value**
LIPC	HDL	30	0.85	3	0.81	57	0.18
LPL	HDL	20	0.54	4	0.13	12	0.81
ANGPTL4	HDL	11	0.17	2	0.98	5	0.02
LIPG	HDL	**16**	**0.02**	2	0.89	**11**	**0.02**
HNF4A	HDL	23	0.15	5	0.28	29	0.80
CD300LG	HDL	15	0.11	**3**	**0.02**	4	0.37
PCSK9	LDL	21	0.21	3	0.09	11	0.82
BCAM	LDL	36	0.28	7	**0.04**	**5**	**0.02**
CBLC	LDL	16	0.86	5	0.77	3	0.39
PVR	LDL	**12**	**0.02**	2	0.08	8	0.12
LDLR	LDL	43	0.93	10	0.94	11	0.29
ANGPTL4	TG	11	0.25	**2**	**0.03**	**5**	**0.05**
LPL	TG	20	0.83	4	0.51	12	0.56

nVars = number of variants analysed, HDL = High Density Lipoproteins, LDL = Low Density Lipoproteins, TG = Triglycerides, MAF = Minor Allele Frequency. Results in bold represent p-values $\leq$ 0.05. No multiple testing threshold was applied to the results of this analysis as the purpose was to compare filtering approaches. All p-values were calculated using SKAT.

doi:10.1371/journal.pone.0154181.t001

Table 2. Results of gene-level rare variant association tests using various variant filters (MAF ≤ 1%).

		Nonsynonymous variants		Loss-of-Function variants		CADD variants (C-Score ≥ 15)	
Gene	**Lipid trait**	**nVars**	**P-value**	**nVars**	**P-value**	**nVars**	**P-value**
LIPC	HDL	27	0.32	2	0.40	54	0.37
LPL	HDL	18	0.12	4	0.13	11	0.54
ANGPTL4	HDL	11	0.17	2	0.98	**5**	**0.02**
LIPG	HDL	15	0.07	2	0.89	**10**	**0.05**
HNF4A	HDL	21	0.32	5	0.28	28	0.60
CD300LG	HDL	14	0.10	**3**	**0.02**	3	0.28
PCSK9	LDL	19	0. 70	2	0.72	11	0.82
BCAM	LDL	32	0.16	7	**0.04**	**3**	**0.02**
CBLC	LDL	12	0.57	4	0.43	2	0.26
PVR	LDL	**11**	**0.02**	1	N/A	8	0.13
LDLR	LDL	41	0.92	9	0.93	11	0.28
ANGPTL4	TG	11	0.25	**2**	**0.03**	**5**	**0.05**
LPL	TG	18	0.45	4	0.51	11	0.26

nVars = number of variants analysed, HDL = High Density Lipoproteins, LDL = Low Density Lipoproteins, TG = Triglycerides, MAF = Minor Allele Frequency. Results in bold represent p-values ≤ 0.05. No multiple testing threshold was applied to the results of this analysis as the purpose was to compare filtering approaches. All p-values were calculated using SKAT.

doi:10.1371/journal.pone.0154181.t002

Gene-based p-values from the CADD filtered analyses were matched with p-values from the 'nonsynonymous' filtered analysis and the results were–log10 transformed and plotted for each trait and MAF cutoff. This meant that only genes which had at least 2 variants within their region after filtering in both sets of analyses (i.e. ≥ 2 'nonsynonymous' & ≥ 2 variants with a CADD C-Score ≥15) were plotted on these graphs. Due to the frequency of points on these plots, overall trends would have been very challenging to identify using scatter plots. We therefore used hexbin plots for this task [18], which allows density of the number of points within each region of the plot to be incorporated. Overall there was little evidence that filtering according to CADD annotations provided either stronger or weaker evidence of association across the genome in comparison filtering by nonsynonymous consequence. This can be observed in the hexbin plots as the gradient of colour is consistent through the plots and does not favour either axis. Moreover, some of the lowest gene-based p-values were only observed using one filtering approach for each trait, implying that there was not always strong concordance between the different methods (i.e. evidence of association was only observed when using CADD filtering and vice versa). Figs 1 and 2 show the hexbin plots for the 4 lipid traits investigated in this analysis for low frequency (MAF≤5%) and rare variant (MAF≤1%) analyses respectively. Plots for the other cardiovascular traits can be found in the S1 File.

Gene-based p-values from all CADD filtered analyses were also matched with those observed from the loss-of-function analyses. However, simple scatter plots were sufficient for this data as there were far fewer than in the comparison with nonsynonymous filtering. It is also worth clarifying that these loss-of-function variants may also have been included in the nonsynonymous analysis, as variants can be classed as both (i.e. a variant which alters the amino acid sequence of a protein and also results in a stop codon). As before, there was no overall trend to suggest that filtering using either approach provided stronger evidence of association with cardiovascular traits. Again there were results which represented a lack of

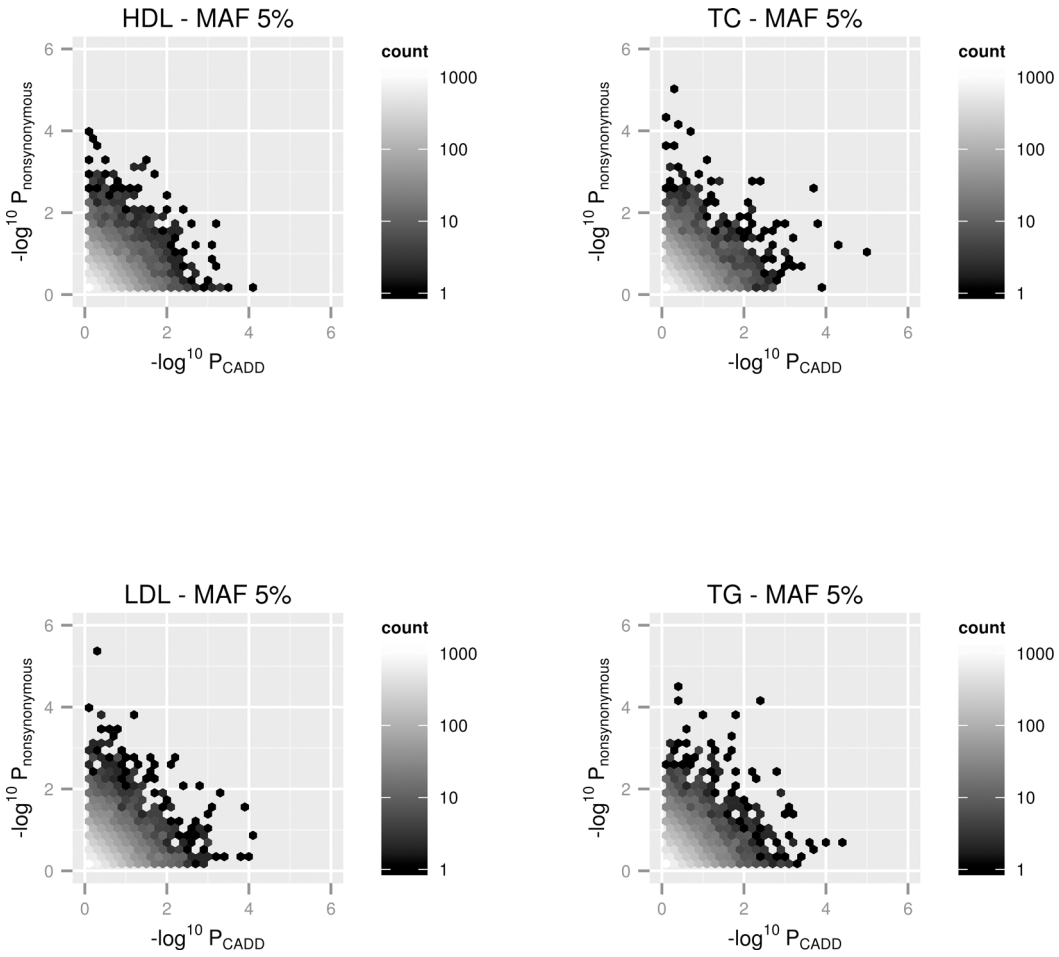

Fig 1. Hexbin plots representing gene-based SKAT analyses for all genes across the genome using a MAF cutoff of 5% with 4 lipid traits. The x-axis represents the–log10 transformed p-value from the analysis after filtering according to CADD annotations. The y-axis represents the–log10 transformed p-value from the analysis after filtering according to 'nonsynonymous' annotations. Only gene regions which had at least 2 variants in them after filtering by both methods were plotted.

doi:10.1371/journal.pone.0154181.g001

concordance between filtering approaches, suggesting that evidence of association with traits would only be observed when filtering with one method but not the other. Scatter plots for these analyses can be found in S1 File. Plots were also generated for the results of identical analyses except using the SKAT-O [19] and MiST [20] collapsing methods, although the overall findings were the same using these approaches. These plots can also be found in S1 File.

Genome-wide Analysis of Gene-based Association Signals using FATHMM-MKL and DANN

The previous analysis was repeated except using two alternative bioinformatics tools rather than CADD. These were FATHMM-MKL [21] and DANN [22], which use a machine learning and deep neural network approach respectively to assess the functional consequence of variants. The metrics for both tools are bounded between 0 and 1, where a larger score means that variants are predicted to have a more deleterious effect. Applying a threshold of 0.9 for both tools (i.e. only variants predicted to have a strong deleterious effect) left 1,208,786 variants using FATHMM-MKL and 1,331,967 variants using DANN in our dataset.

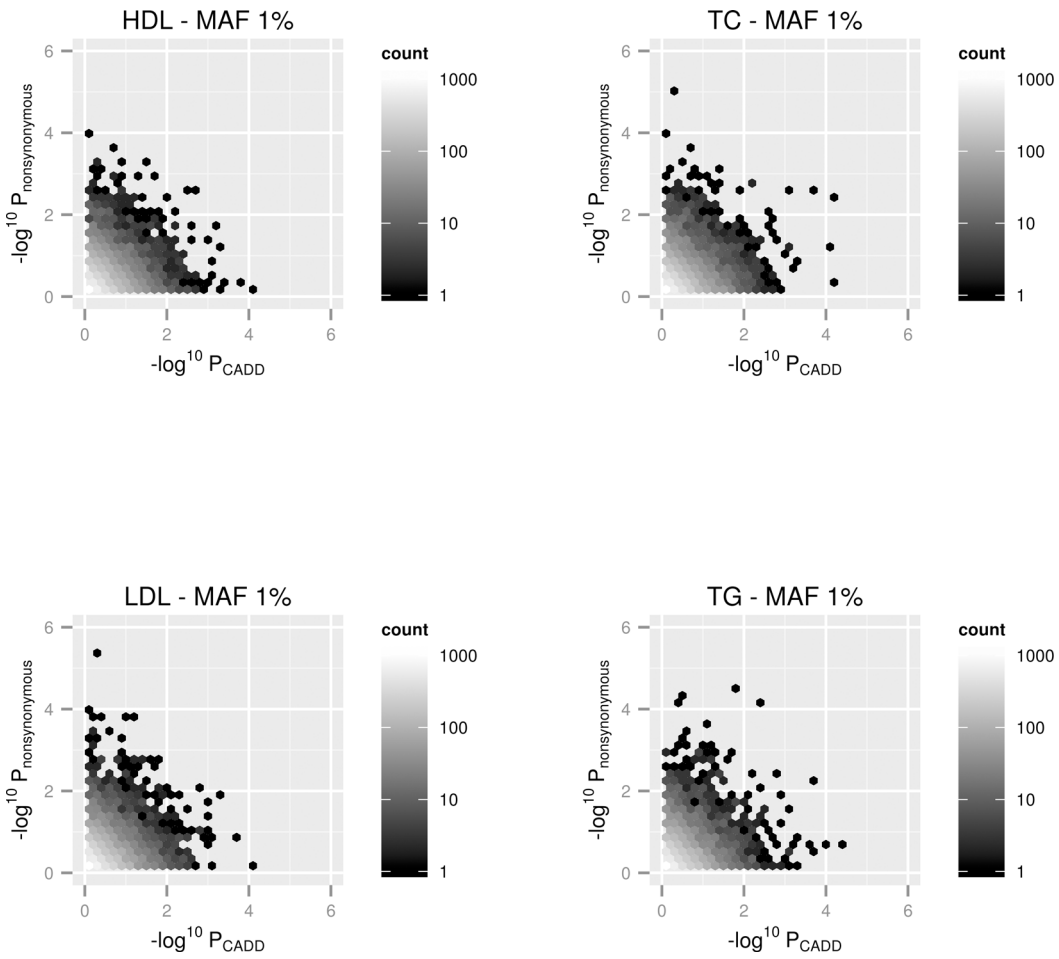

Fig 2. Hexbin plots representing gene-based SKAT analyses for all genes across the genome using a MAF cutoff of 1% with 4 lipid traits. The x-axis represents the–log10 transformed p-value from the analysis after filtering according to CADD annotations. The y-axis represents the–log10 transformed p-value from the analysis after filtering according to 'nonsynonymous' annotations. Only gene regions which had at least 2 variants in them after filtering by both methods were plotted.

doi:10.1371/journal.pone.0154181.g002

FATHMM-MKL and DANN analyses were conducted separately. Analyses were undertaken as before, collapsing variants by gene regions and analysing them with each cardiovascular trait in turn after applying a MAF cutoff of 5% and 1%. Gene regions were analysed twice using the SKAT-O and MiST tests and only regions which had at least 2 variants after filtering were eligible for analysis. Hexbin and scatter plots were again used to display results compared to the 'nonsynonymous' results and the 'loss-of-function' results respectively. The FATHMM-MKL plots suggested similar inferences to the CADD based analyses, whereas certain DANN analyses using SKAT-O (particularly for TG) appeared to have more concordance with results using nonsynonymous annotations. Plots from all these analyses can be found in S1 File.

Discussion

We have conducted a candidate gene study to evaluate whether incorporating variant annotations using a non-coding bioinformatics tool can aid rare variant analyses over filtering by variant consequence. Filtering according to CADD provided evidence of association between genes

and lipids similar to filtering by variant type, as well as moderate association between *ANGPTL4* and HDL (P = 0.02). We have also undertaken extensive genome wide analyses to determine how filtering variants using bioinformatics annotations compares to filtering by variant consequence within gene regions, when conducting low frequency and rare variant collapsing analyses. These results were plotted as hexbin and scatter plots, which suggested the different approaches to filtering variants yield different sets of associated loci. The consequence of this meant that certain association signals were only detected when filtering by annotations from the bioinformatics tools and not according to variant consequence, and vice versa. This suggests that future studies should benefit from this alternative approach to identify association signals potentially overlooked using conventional methods, due to the inclusion of non-coding regions which may harbour potentially deleterious variants.

The reason why this novel approach to filtering variants may identify association signals not detected when filtering by variant consequence could be influenced by several factors. Firstly, this could be due to the inclusion of predicted deleterious synonymous variants which therefore improves the statistical power of analyses. Furthermore, using these annotations may lead to removing predicted neutral nonsynonymous variants (i.e. variants which alter amino acid sequence, but not the function of a protein) which reduces the amount of statistical noise incorporated into analyses. Fig 1B of the CADD manuscript by Kircher et al shows that a threshold of 15 should also be including a proportion of other types of variants (e.g. which are predicted to be the most deleterious within each of their categories, along with the majority of nonsynonymous variants). Filtering using a non-coding algorithm also allows the inclusion of variants which reside in intronic regions of genes, which can also have functional consequences on human disease through altering regulator or splicing sequence [23]. Association signals driven by variants within these regions may therefore be of great importance in terms of disease aetiology, although collapsing approaches which filter by variant consequence are confined to coding regions of the genome and therefore these variants are not investigated. Incorporating non-coding bioinformatics annotations into rare variant analysis should therefore aid future studies in terms of addressing the limitations of conventional approaches.

The ability to adjust the inclusion threshold for variants when using prediction algorithm annotations is also advantageous to studies. In this study, we have used a CADD C-Score of $\geq$ 15 as our inclusion threshold, which is suggested by the authors of CADD to identify potentially pathogenic variants as it is the median value for all possible canonical splice site changes and nonsynonymous variants [9]. However, the optimal threshold of CADD C-Score to uncover causal variants may depend on several factors, such as the trait analysed. With the particular interest regarding the contribution that low frequency and rare variants can have on blood lipid levels currently [17, 24, 25], we decided to analyse lipid and other cardiovascular traits in this study to evaluate our hypothesis. However, filtering using bioinformatics annotations should also benefit analyses undertaken with other complex traits. Likewise, the length of a gene is another factor which may cause varying cutoffs to result in stronger evidence of association detected from analyses. It may therefore be beneficial to use a stricter threshold (i.e. more confidence that variants within a region are deleterious) for larger gene regions, although this will likely depend on the hypothesis of the study.

In this study we have predominantly used the SKAT-O and MiST tests as they have been reported to be amongst the most consistently powerful collapsing methods according to evaluations of gene-based tests [26]. However, regardless of the choice over collapsing method when undertaking a rare variant analysis, the filtering phase is crucial to identifying association signals from causal variants. Although the 3 bioinformatics tools used in this study all have a similar purpose, the manner in which they accomplish this is quite varied. To predict the effect of non-coding variants, CADD uses a support vector machine with a linear kernel,

FATHMM-MKL uses multiple kernel learning and DANN uses a deep neural network. Moreover, even though CADD and DANN use the same training data for their tools, FATHMM-MKL uses pathogenic data from the Human Gene Mutation database [27] and control data from the 1000 genomes project [28]. As a result, the correlation between the annotations from these tools is not that strong, which is why they have all been utilised in this study (CADD & FATHMM-MKL = 0.62, CADD & DANN = 0.74 and DANN & FATHMM-MKL = 0.56 according to dbNSFP v3.0 [29]).

Despite the advances in low coverage sequencing techniques over the last few years, the majority of rare variant analyses applied to these data have been underwhelming[30, 31]. Recently, studies have attempted to improve statistical power in their approaches by increasing their sample sizes[17], adapting current methodological techniques[32] or conducting analyses on remote or isolated populations where the allele frequency of rarer variants may be heightened[33]. We suggest that incorporating predicted functionality of variants into analyses (based on the wealth of functional annotation data now available) should prove to be a valuable and feasible addition to these options in identification of association signals for future studies. Moreover, there has been a greater emphasis recently on elucidating the functional role of variants[34–36], for which filtering approaches using resources such as the bioinformatics tools utilised in this study could be integrated to great effect.

The fundamental question when filtering variants according to predicted function concerns the reliability of the bioinformatics resource used. Clearly, the success of any analysis which relies on these resources hinges on their accuracy. Further advances in the accuracy of variant prediction will have a beneficial impact on association studies which incorporate annotations as we have illustrated in this study. Moreover, filtering variants based on prediction algorithms can be undertaken for analyses conducted in non-coding regions of the genome, which would not be possible when filtering according to variant consequence. The majority of GWAS hits discovered to date fall within non-coding regions[37, 38], although rare variant analyses using collapsing approaches have so far been confined to coding regions of the genome. Therefore, filtering according to non-coding prediction algorithms provides a platform for future studies to investigate the role of variants in these regions, such as investigating the impact of variants in flanking regions of genes which may have an impact on regulatory variation from RNA to protein. This should be of particular interest to future studies given the amount of WGS data currently in development.

Conclusion

Filtering low frequency and rare variants using knowledge based on molecular function and pathogenicity should help identify strong evidence of association not detected using conventional filtering approaches. Follow up analyses which evaluate these signals will be beneficial in the identification of potential mechanisms and causal variants for complex disease.

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 (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[39, 40] (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[41]. 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)[42]. 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[43] 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[44] to exclude participants of non-European ancestry after merging our data with a pruned 11 HapMap3 population dataset[45]. 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 (1,754 TwinsUK and 1,867 ALSPAC).

Phenotype data

ALSPAC. Height was measured to the nearest 0.1cm using a Harpenden stadiometer (Holtain Crosswell, Dyfed, UK) and weight was measured to the nearest 0.1kg using Tanita electronic scales. Body Mass Index (BMI) was calculated as (weight (kg))/(height (m))2. Blood Pressure was measured with a Dinamap 9301 vital monitor completed by trained staff using the appropriate cuff size. Two readings of both systolic and diastolic blood pressure (SBP &

DBP respectively) were taken when the study participants were at rest and the mean of each were used as a measurement in our analysis. Both these measurements were taken from the age 9 clinic (mean age: 9.9, range: 8.9–11.5).

Non-fasting blood samples were also 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 (HDL)) were measured by modification of the standard Lipid Research Clinics Protocol with enzymatic reagents for lipid determination [46]. Low density lipoprotein cholesterol (LDL) concentration was subsequently calculated using the Friedwald equation[47]:

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

TwinsUK. Height was measured to the nearest 0.5cm using a wall-mounted stadiometer and weight (light clothing only) was measured to the nearest 0.1kg using digital scales. Body Mass Index (BMI) was calculated as (weight (kg))/(height (m))^2. Brachial blood pressure was measured using an automated cuff sphygmomanometer (OMRON HEM713C; Omron Healthcare (UK) Ltd, Henfield, UK). SBP and DBP were measured three times, two of which were highly correlated (0.90 for SBP and 0.92 for DBP) and averaged to get our final phenotype measurements.

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[48]. A colorimetric enzymatic method was used to determine TC, TG and HDL levels. The Friedewald equation was used to calculate LDL levels in subjects.

Statistical Analysis

We used CADD[9] to obtain scaled C scores for all 44.9 million possible variants and indels in the UK10K whole genome sequence data[14]. After removing variants which failed QC, we filtered all variants in three ways 1) responsible for a 'nonsynonymous' substitution according to the Variant Effect Predictor[15] (VEP) 2) responsible for a 'loss of function' according to VEP (i.e. 'stop losses/gains', 'splice sites' or 'frameshift indels') and 3) Variants with a CADD C-Score of 15 or higher. We collapsed variants together across candidate genes and analysed them using SKAT[8] with their associated traits according to Liu et al[17]. These genes were *ANGPTL4*, *BCAM*, *CBLC*, *CD300LG*, *HNF4A*, *LDLR*, *LIPC*, *LIPG*, *LPL*, *PCSK9* and *PVR*. Analyses were repeated after applying each variant filtering method, as well as applying two different minor allele frequency cutoffs (MAF$\leq$5%) and (MAF$\leq$1%).

We also identified variants with a CADD C-Score $\geq$15 across the genome and aggregated them together across all gene regions according to UCSC definitions (reference genome hg19). Variants were then collapsed together and analysed with 7 cardiovascular traits (BMI, SBP, DBP, TC, HDL, LDL & TG) using SKAT after applying a MAF cutoff of 5%. We repeated this process in a second set of analyses using a MAF cutoff of 1% to investigate rarer variation.

Gene-based p-values from the CADD analyses were matched with p-values from the 'nonsynonymous' filtered analysis and the results were–log10 transformed and plotted using hexbin plots for each trait and MAF cutoff. This meant that only genes which had at least 2 variants within their region after filtering in both sets of analyses (i.e. $\geq$ 2 'nonsynonymous' & $\geq$ 2 variants with a CADD C-Score $\geq$15) were plotted on these graphs. This was undertaken using the 'hexbin' package in R [18] This process was repeated except matching on results from the 'loss-

of-function' analysis. As this resulted in far fewer data points to be plotted, scatter plots were generated instead using using the R package 'ggplot2' [49]. All plots were also regenerated using the results of an identical analysis except using the SKAT-O [19] and MiST tests [20]. This approach was repeated except using two alternative bioinformatics tools to CADD which can also predict the effect of variants in both coding and non-coding regions of the genome (FATHMM-MKL [21] and DANN [22]). R statistical software[50] was used for all statistical analyses and plots.

Supporting Information

S1 File. Gene based low frequency and rare variant analyses using various variant filtering approaches and collapsing methods.
(PDF)

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. 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 IEU MC_UU_12013/8). TGR is a UK MRC PhD Student.

Author Contributions

Conceived and designed the experiments: TR TG. Analyzed the data: TR. Wrote the paper: TR TG. Made valuable contributions to the manuscript with suggested changes: CC NT.

References

1. 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 PMID: 20479773.
2. Wagner MJ. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. Pharmacogenomics. 2013; 14(4):413–24. doi: 10.2217/pgs.13.36 PMID: 23438888.
3. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annual review of genetics. 2010; 44:293–308. doi: 10.1146/annurev-genet-102209-163421 PMID: 21047260.
4. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456(7218):18–21. doi: 10.1038/456018a PMID: 18987709.
5. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008; 40(6):695–701. doi: 10.1038/ng.f.136 PMID: 18509313; PubMed Central PMCID: PMC2527050.
6. Sul JH, Han B, Eskin E. Increasing power of groupwise association test with likelihood ratio test. Journal of computational biology: a journal of computational molecular cell biology. 2011; 18(11):1611–24. doi: 10.1089/cmb.2011.0161 PMID: 21919745; PubMed Central PMCID: PMC3216097.
7. 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 PMID: 21408211; PubMed Central PMCID: PMC3048375.

8. 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 PMID: 21737059; PubMed Central PMCID: PMC3135811.
9. 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(3):310–5. doi: 10.1038/ng.2892 PMID: 24487276; PubMed Central PMCID: PMC3992975.
10. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003; 31(13):3812–4. PMID: 12824425; PubMed Central PMCID: PMC168916.
11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4):248–9. doi: 10.1038/nmeth0410-248 PMID: 20354512; PubMed Central PMCID: PMC2855889.
12. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014; 95(1):5–23. doi: 10.1016/j.ajhg.2014.06.009 PMID: 24995866; PubMed Central PMCID: PMC4085641.
13. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(17):6131–8. doi: 10.1073/pnas.1318948111 PMID: 24753594; PubMed Central PMCID: PMC4035993.
14. Muddyman D, Smee C, Griffin H, Kaye J. Implementing a successful data-management framework: the UK10K managed access model. Genome medicine. 2013; 5(11):100. doi: 10.1186/gm504 PMID: 24229443; PubMed Central PMCID: PMC3978569.
15. 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(16):2069–70. doi: 10.1093/bioinformatics/btq330 PMID: 20562413; PubMed Central PMCID: PMC2916720.
16. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–8. doi: 10.1093/bioinformatics/btr330 PMID: 21653522; PubMed Central PMCID: PMC3137218.
17. 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 PMID: 24336170; PubMed Central PMCID: PMC3939031.
18. Lewin-Koh N. Hexagon binning: an overview, Technical Report. http://cranr-projectorg/web/packages/hexbin/vignettes/hexagon-binningpdf. 2011.
19. 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 PMID: 22863193; PubMed Central PMCID: PMC3415556.
20. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. Genetic epidemiology. 2013; 37(4):334–44. doi: 10.1002/gepi.21717 PMID: 23483651; PubMed Central PMCID: PMCPMC3740585.
21. 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 PMID: 25583119.
22. 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 PMID: 25338716; PubMed Central PMCID: PMCPMC4341060.
23. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science. 2008; 322(5903):881–8. doi: 10.1126/science.1156409 PMID: 18988837; PubMed Central PMCID: PMC2694957.
24. Timpson NJ, Walter K, Min JL, Tachmazidou I, Malerba G, Shin S- Y, et al. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. Nature communications. 2014;5. doi: 10.1038/ncomms5871
25. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American journal of human genetics. 2014; 94(2):223–32. doi: 10.1016/j.ajhg.2014.01.009 PMID: 24507774; PubMed Central PMCID: PMC3928662.
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 PMID: 25906071; PubMed Central PMCID: PMC4407972.

27. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics. 2014; 133(1):1–9. doi: 10.1007/s00439-013-1358-4 PMID: 24077912; PubMed Central PMCID: PMC3898141.

28. 1000 Genomes Project, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. doi: 10.1038/nature11632 PMID: 23128226; PubMed Central PMCID: PMC3498066.

29. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human mutation. 2016; 37(3):235–41. doi: 10.1002/humu.22932 PMID: 26555599; PubMed Central PMCID: PMCPMC4752381.

30. 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. European journal of human genetics: EJHG. 2013; 21(9):1027–30. doi: 10.1038/ejhg.2012.284 PMID: 23321613; PubMed Central PMCID: PMC3746260.

31. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature. 2013; 498(7453):232–5. doi: 10.1038/nature12170 PMID: 23698362; PubMed Central PMCID: PMC3736321.

32. Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genetic epidemiology. 2013; 37(5):409–18. doi: 10.1002/gepi.21727 PMID: 23650101; PubMed Central PMCID: PMC3706099.

33. Tachmazidou I, Dedoussis G, Southam L, Farmaki AE, Ritchie GR, Xifara DK, et al. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. Nature communications. 2013; 4:2872. doi: 10.1038/ncomms3872 PMID: 24343240; PubMed Central PMCID: PMC3905724.

34. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014; 507(7492):371–5. doi: 10.1038/nature13138 PMID: 24646999; PubMed Central PMCID: PMC4113484.

35. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. The New England journal of medicine. 2013; 368(2):117–27. doi: 10.1056/NEJMoa1211851 PMID: 23150934; PubMed Central PMCID: PMC3631573.

36. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493(7431):216–20. Available: http://www.nature.com/nature/journal/v493/n7431/abs/nature11690.html#supplementary-information. doi: 10.1038/nature11690 PMID: 23201682

37. Manolio TA. Genomewide association studies and assessment of the risk of disease. The New England journal of medicine. 2010; 363(2):166–76. doi: 10.1056/NEJMra0905980 PMID: 20647212.

38. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(23):9362–7. doi: 10.1073/pnas.0903103106 PMID: 19474294; PubMed Central PMCID: PMC2687147.

39. 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.

40. 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 PMID: 22507743; PubMed Central PMCID: PMC3600618.

41. 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 PMID: 23088889; PubMed Central PMCID: PMC3927054.

42. 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 PMID: 20080505; PubMed Central PMCID: PMC2828108.

43. 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 PMID: 17924348; PubMed Central PMCID: PMC2265661.

44. 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 PMID: 16862161.

45. 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 PMID: 20811451; PubMed Central PMCID: PMC3173859.

46. 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.

47. Warnick GR. Laboratory measurement of lipid and lipoprotein risk factors. Scandinavian journal of clinical and laboratory investigation Supplementum. 1990; 198:9–19. PMID: 2189213.

48. 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 PMID: 20388742.

49. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed: Springer Publishing Company; 2009.

50. Team RDC. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria2008.