The contribution of rare whole genome sequencing variants to plasma protein levels and to the missing heritability

Marcin Kierczak
Uppsala University https://orcid.org/0000-0003-2629-5655
Nima Rafati
Uppsala University
Julia Höglund
Uppsala University https://orcid.org/0000-0001-8061-3947
Hadrien Gourle
Uppsala University https://orcid.org/0000-0001-9807-1082
Daniel Schmitz
Uppsala University https://orcid.org/0000-0003-4480-891X
Weronica Ek
Uppsala University https://orcid.org/0000-0003-2194-496X
Stefan Enroth
Uppsala University https://orcid.org/0000-0002-5056-9137
Diana Ekman
Science for Life Laboratory
Björn Nystedt
Stockholm University https://orcid.org/0000-0001-7809-7664
Torgny Karlsson
Uppsala University https://orcid.org/0000-0001-8095-6149
Åsa Johansson (asa.johansson@igp.uu.se)
Uppsala University https://orcid.org/0000-0002-2915-4498

Article

Keywords: Rare variants, SKAT, Protein Biomarkers, Hidden heritability, Missing heritability, GWAS

Posted Date: June 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-625433/v1

Version of Record: A version of this preprint was published at Nature Communications on May 9th, 2022.
See the published version at https://doi.org/10.1038/s41467-022-30208-8.

Abstract

Despite the success in identifying effects of common genetic variants, using genome-wide association studies (GWAS), much of the genetic contribution to complex traits remains unexplained. Here, we analysed high coverage whole-genome sequencing (WGS) data, to evaluate the contribution of rare genetic variants to 414 plasma proteins. The frequency distribution of genetic variants was skewed towards the rare spectrum, and damaging variants were more often rare. However, only 2.24% of the heritability was estimated to be explained by rare variants. A gene-based approach, developed to also capture the effect of rare variants, identified associations for 249 of the proteins, which was 25% more as compared to a GWAS. Out of those, 24 associations were driven by rare variants, clearly highlighting the capacity of aggregated tests and WGS data. We conclude that, while many rare variants have considerable phenotypic effects, their contribution to the missing heritability is limited by their low frequencies.

Introduction

Since the advent of genome-wide association studies (GWAS), thousands of genetic variants have been identified to be associated with common diseases and other health-related traits. However, the genetic variants identified to date explain a limited part of the heritability of most common diseases and traits. For example, in one of the largest GWAS for body mass index (BMI) until now, 941 significant lead SNPs explain only 6% of the variation[1]. However, by combining the effects of all SNPs together, regardless of whether they reach the genome-wide significance threshold, one can capture a substantial part of the heritability of a trait[2], i.e., SNP heritability. Here, our working hypothesis is that a part of the complex traits' heritability can be attributed to the effect of rare genetic variants. These rare variants are, due to only being present in one or few individuals of a cohort, not identified as genome-wide significant in GWAS due to lack of power. However, when analysing the aggregated effect of several rare variants in a region, their effects are more likely to be captured.

One limitation in previous studies is that the vast majority of GWAS has been performed using data originating from SNP arrays, where rare variants are substantially underrepresented. Whole-genome sequencing (WGS) is soon to become a golden standard in large-scale genetic studies, allowing for easy characterisation of genetic variations, such as single nucleotide variants (SNVs) and short insertions and deletions (indels) at any frequency. In light of these advances, GWAS for complex diseases are no longer limited to analysing only the common variants. In natural populations, purifying selection acts against deleterious variants and keeps the frequency of such variants low. Rare variants are therefore more likely to be functionally important than common variants. One example is the melanocortin 4 receptor gene (*MC4R)*, where one common allele is associated with a small (~ 0.25 kg/m^2) increase in BMI[3]. However, very rare deleterious mutations in *MC4R* represent the most common monogenic cause of severe early onset obesity[4]. Variants that have such considerable effects on a trait are often rare or even specific to individual populations or families and are therefore impossible to detect using SNP genotyping data.

Thus, WGS data is required to investigate the effect of rare variants on human traits. By performing WGS on more than 1000 Swedish samples, we have recently shown that rare variants constitute the major part of all genetic variants in a cohort[5,6]. This supports the need to also analyse the effect of rare variants in relation to common diseases and traits.

The low frequency of most sequence variants[5] seriously limits the power of using a GWAS strategy (single-marker tests) for analysing WGS data[6]. In order to overcome these limitations, rare variants can be collapsed and analysed jointly in a burden test[7–9]. Burden tests are powerful when the effects of all variants collapsed have similar magnitudes and directions, for example when two different missense variants influence the protein function to the same degree and are both disadvantageous rather than one being disadvantageous and one being beneficial. By filtering on the predicted deleteriousness of rare variants, and including only loss of function (LOF), or possibly deleterious variants, various studies have shown that the burden of damaging rare variants is associated with complex traits and diseases[10,11]. In addition to filtering, variants can also be weighted by their allele frequency, assuming that rare variants are more pathogenic than the common ones, or be weighted by their predicted deleteriousness[12]. However, while rare variants are more likely to have a functional consequence than common variants, most genetic variants in the genome, including the rare ones, are neutral[13]. Therefore, kernel association methods that do not assume uniform directionality and magnitude of all tested variables, are more appropriate to use when analysing WGS data. One such method, the Sequence Kernel Association Test (SKAT)[12], modulates the effect of multiple genetic variants together in a multivariable approach. In SKAT, each variant can be assigned a weight that reflects its expected impact on the test statistics. Commonly, either variants annotated as damaging are upweighted, or rare variants are upweighted since they are more likely to be deleterious and the power to detect the effect of a rare variant is typically limited by a small number of observations.

The majority of studies aiming to investigate the effect of rare variants share the common approach of using a test that collapses the effect of multiple genetic variants. It is also common that a cut-off for minor allele frequency (MAF)[11,14] is used, sometimes limiting the analysis to only very (ultra) rare[10] variants, for example, only observed in one or a few individuals, or low frequency (MAF < 5%) variants[14]. Here, we considered a MAF threshold of $1/\sqrt{(2 * \text{sample size})}$, which equals 2.39% in our study, between common and rare variants as this has been well-motivated in a previous study, and is the default value in the SKAT model[15]. However, it is clear from previous GWAS that common variants with small effects play an important role in the development of common diseases. Our hypothesis is that we can, in addition to analysing common and rare variants separately, gain even more power from incorporating information from both common and rare variants in an aggregated test. As the power to identify the effect of common variants is much larger, not only in a GWAS, but also in a multivariate approach, the effect of rare variants can still be hard to capture. However, analysing common and rare variants separately and combining the test statistics, increases the power to identify the effect of both common and rare variants simultaneously[15].

In this study, we have analysed SNVs and indels, identified by deep coverage WGS, in relation to the protein expression levels of 414 plasma proteins (Supplementary Table S1), either well known or exploratory biomarkers for disease, using the SKAT method. Our aim was to identify the contribution of rare genetic variants, on top of the common variants that can be identified using a standard GWAS approach. In previous studies, we have performed GWAS on the protein abundance levels of the same proteins using genotyped and imputed SNPs. We have shown that the levels of many of these protein have high heritability, with up to 67%[16], and are strongly influenced by genetic variants commonly located in the regulatory regions of the gene encoding the protein itself. The protein dataset is therefore very well suited for investigating not only the effect of rare coding variants, but also the effect of regulatory non-coding variants. In this study we analysed the same proteins as in our previous GWAS[6,16–19], but here we are using WGS data in combination with gene-based kernel association tests to also include and identify the contribution of rare variants to protein level variation in plasma protein levels.

Results

In total, 872 participants passed WGS and protein QC (Supplementary Figure S1) out of which 443 (50.8%) were females. The ages ranged between 14 and 94 (median 50) years. In total, 16,271,782 variants had been called of which 12,956,981 biallelic SNVs and 1,130,297 biallelic indels passed QC. By using a MAF threshold equal to $1/\sqrt{(2 * \text{sample size})} = 2.39\%$ as the upper limit for considering a variant to be rare, in agreement with previous suggestions[15], nearly half (49.4%) of the variants were considered rare. There was a clear skew towards the lower frequency spectrum (Fig. 1A).

Common variants explain most of the heritability

Even if a considerable fraction of identified variants were considered rare in the cohort (Fig. 1A), each individual carries a considerably larger number of common alleles (Fig. 1B). On average, only 2.27 % (standard deviation, S = 0.63%) of the variants in each individual were considered being rare in the cohort (dark grey in Fig. 1B). However, there was clearly (χ^2 P-value < 2.2×10^{-16}) a much larger fraction of rare variants among the high CADD and high Eigen values (Fig. 1C). For example, while 50% the variants with low CADD values (< 10) were rare, as many as 95% of the variants with the strongest predicted deleterious effects (CADD > 40) were rare. This is in line with the idea that rare variants are more likely to have deleterious effects. However, since each individual has a much larger number of alleles that are common in the cohort (Fig. 1B), the total per-individual burden of potentially damaging variants, as predicted by CADD (Fig. 1D) or Eigen (Supplementary Figure S2), is mainly due to common variants (P-value for trend < 2.2×10^{-16}). Using our definition of rare (MAF < 2.39%), resulted in that only a minor fraction of the total per-individual burden of damaging effects could be attributed to the rare variants: 2.24 % (SD = 0.58%) or 2.24% (SD = 0.57%) when CADD and Eigen values were used to predict the deleteriousness. Since the heritability is the amount of variation in a phenotype that is due to genetic effects, the expected fraction of the heritability explained by rare variants will be the same as the average per-individual phenotypic effects in the cohort, indicating that a very limited amount of the heritability is indeed due to rare variants.

This suggests that a larger fraction (estimated to 97.76%) of the total amount of the heritability can be attributed to common variants.

GWAS results

As a comparison to the results from our SKAT analyses, we first performed a traditional GWAS for each protein. However, to include as many lead GWAS SNVs and indels as possible in the conditional SKAT analyses, we first used a liberal significance threshold of $5x10^{-8}$ in the GWAS. A total of 274 proteins had at least one significant SNV or indel (217 proteins had any *Cis* and 107 any *Trans*) association (Supplementary Table S2). The lead variants from the GWAS were skewed towards variants with lower MAF (Fig. 2A), which agrees with, the MAF distribution in the cohort, but is less pronounced (Fig. 1A). However, the power to detect an association drops dramatically for rare variants down to a count of three alleles in the population, where the power is zero[6]. Among the GWAS-identified variants, the rare variants tend to have larger effect sizes (Fig. 2B) than the common ones, with average betas of 1.52 and 0.62 for rare and common respectively ($P < 2.2 \times 10^{-16}$). Using a stricter P-value cut-off for *Trans* regulatory effects, $3.92x10^{-11}$, and $3.00x10^{-8}$ for *Cis*, which would be more appropriate for identifying novel GWAS hits when 414 proteins are analysed, resulted in 235 proteins having at least one significant hit (213 *Cis*, 42 *Trans*).

SKAT analyses

In the SKAT analyses (see Methods section for details), five different types of SNV-sets were constructed (Supplementary Figure S3), and seven different SKAT models, with different weighting and/or filtering of variants based on MAF (Supplementary Figure S4), CADD or Eigen values, were analysed for each SNV-set (Table 1, Fig. 3A). A total number of 249 (61%), out of the 405 proteins encoded by genes located on autosomal chromosomes, were associated (5.88×10^{-6}) with at least one of the *Cis*-SNV-sets (Fig. 3A, Table 1, Supplementary Table S3). Among those, 208 had a significant association with a *Cis*-Reg-set, 205 with a *Cis*-Flank-set, and 195 with a *Cis*-CDS-set. For all SNV-sets, the largest number of significant protein associations was identified using the CommonRare method (Model 6) in SKAT, and the lowest number was identified when only analysing rare variants (Model 7) closely followed by the model (Model 3) where rare variants were highly upweighted (Table 1, Fig. 3B). It is also worth highlighting that, for any of the SNV-sets, *Cis* SKAT analyses, weighted by Eigen scores (Model 2), did not appear to result in a larger number of significant results compared to no weighting at all (Model 1), for any of the SNV-sets (Table 1).

Table 1
Overview and summary statistics for the five types of SNV-sets analysed and the number of significant findings obtained with each of the seven SKAT models used.

Cis/ *Trans*	SNV-set	No SNV-sets[a]	Number of SNVs[b]	Rare[c]	No. proteins with significant associations for the different models[d]						
					1	2	3	4	5	6	7
Cis	*Cis*-Reg	405	61 [31–116]	47.0%	158	154 [g]	83	161	175	194	47
Cis	*Cis*-Flank[e]	405	190 [118–307]	48.4%	163	157 [g]	74	157	171	198	45
Cis	*Cis*-CDS	405	8 [5–13]	49.1.%	138	132 [h]	87	150	158	181	53
Trans	*Trans*-CDS[f]	18,467	8 [5–14]	54.2%	18	23 [h]	6	18	21	44	11
Trans	*Trans*-Flank	18,467	229 [161–330]	51.3%	18	19 [g,h]	8	19	24	61	26

a) *For Cis-SNV-sets, each of the 405 autosomal SNV-sets were analysed only in relation to the encoded protein, whereas in the Trans-SNV-sets, the SNV-set (one for each of the 18,467 genes across the genome) was analysed in relation to all 414 proteins. The significance threshold was 0.05 / 3 Cis-sets / 405 proteins / 7 models = 5.88e-06 for Cis, and 0.05 / 2 Trans-sets /414 proteins/18,467 SNV-sets / seven models = $4.67x10^{-10}$ for Trans*
b) Median [interquartile range] of the number of SNVs in the SNV-sets
c) Fraction of SNVs and indels in the SNV-sets that were considered rare (MAF < 0.0239)

d) The seven models are: Model 1) Unweighted, Model 2) CADD or Eigen weighted, Model 3) MAF weighted - β(1, 25), Model 4) MAF weighted – β(1, 5), Model 5) MAF weighted – β(0.5, 0.5), Model 6) CommonRare, Model 7) Rare only. See method section for more information on the models and Supplementary Figure S4 for information on the β-distributions

e) ± 100kb gene-regions up/down-stream of each gene, filtered by Eigen > 10 when analysed in Cis
f) ± 100kb gene-regions up/down-stream of each gene, filtered by CADD or Eigen > 10 when analysed in Trans
g) Weighted by Eigen values

h) Weighted by CADD values

Rare variants driving Cis-associations.

The overlap between the GWAS and the SKAT analyses was very large, with 203 proteins having significant Cis-associations detected by both methods. The GWAS identified significant Cis-associations with four proteins (CDH6, FGF-5, IL-18BP, and IL-1RA), for which there were no significant results in any of the SKAT-analyses. In contrast, there were 45 proteins with significant results in the SKAT analyses that

were not identified with GWAS (Fig. 4A, Supplementary Table S3). In addition, among the proteins with overlapping GWAS and SKAT hits, 83 proteins were still significantly associated in the SKAT analyses after adjusting for common GWAS hits (Supplementary Table S3). This indicates that these signals might be driven by rare variants which was supported by that 36 of these proteins were significant in using the rare only model (Model 7). However, other signals could also have been driven by multiple (common and/or rare) variants, in addition to the ones that reached genome-wide significance in the GWAS. For THY-1, AGRP, Ep-CAM, AZU1 and CADM3, the most significant SKAT P-values, for any of the SNV-sets, were estimated when applying the rare-only function (Model 7), including only rare variants. Each of these proteins was significantly associated with a rare variant also in the GWAS (Supplementary Table S2), indicating that these signals are mainly driven by one single rare variant. Also, for 56 proteins, the most significant P-value came from a test where rare variants were upweighted (Model 3–5). Of these, as many as 53 proteins were also associated with a common SNV in the GWAS. However, when adjusting for these common GWAS SNVs, 12 proteins (TNFRSF11A, hK11, hOSCAR, GPC1, RETN, TNFRSF4, TR-AP, WISP-1, CST5, CD6, IL-15RA, and SKR3) still remained significant. This suggests that, for a subset of proteins, we capture additional effects from rare variants that cannot be identified with the GWAS. However, since the majority (N = 41) of the 53 proteins were no longer significant when adjusting for the GWAS hits, it appears that the SKAT methods with rare variants upweighted, to some degree, capture the same underlying genetic effects as the GWAS.

Gene-based test performs much better than GWAS for Trans-associations and the overlap between GWAS and SKAT Trans-associations is small.

In total, 76 proteins had a Trans-association (Supplementary Table S4). A significantly larger (χ^2 P = 3.6 x 10^{-7}) fraction of proteins (76 of 414) have a Trans-association (Supplementary Table S2), compared to the GWAS (42 of 414), when considering the stricter cut-off for significance to adjust for multiple testing in the GWAS. This clearly shows an increased power in identifying Trans-associations when using a gene-based test compared to a traditional GWAS strategy. There was a considerably smaller overlap between the SKAT analyses and the GWAS-results for the Trans-associations (Fig. 4B) than for the Cis-associations (Fig. 4A). However, among the 35 Trans-SKAT loci that overlapped with a common GWAS association, only seven remained significantly associated in the SKAT tests after conditioning on the lead GWAS SNVs (Supplementary Table S4).

Pleiotropic associations are commonly observed in Trans.

Among the 76 proteins with significant Trans-hits, several were associated with SNV-sets at multiple loci and/or several neighbouring SNV-sets within the same loci. A total of 171 (Fig. 5) loci represented by 815 significant SNV-set associations were identified (Supplementary Table S4). Several of the identified loci were pleiotropic, i.e., associated with several of the measured proteins (Fig. 5).

Including non-coding regions increases the power to detect associations.

For Trans-SKAT associations there was a larger number of significant hits for the Trans-Flank-sets than the Trans-CDS-sets (Fig. 4A). The CommonRare function in SKAT in combination with analysing the Trans-Flank-sets, resulted in the largest number (N = 495) of tests with significant P-values (Supplementary Table S4), and a larger fraction of proteins with a significant Trans hit (Fig. 4A). This suggests that including SNVs outside of the coding regions in the analyses increases the power to detect associations, probably due to that ± 100kb gene-regions included in the Trans-Flank-sets include also potential regulatory regions in addition to coding variants.

Pinpointing causal genes in Trans, illustrated by the ABO and TNFRSF10C examples.

For some proteins, tens of neighbouring genes were located within each Trans-Flank-set, and several partly overlapping Trans-Flank-sets (including partly the same genes) gave significant P-values. Therefore, the potentially causal genes behind these associations were not easy to determine. For example, there is a strong association in the ABO region on chromosome 9 for several of the proteins (Fig. 5, Supplementary Table S3). The most significant association for the protein CDH5 is for the Trans-Flank-set surrounding OBP2B ($P = 6.86x10^{-48}$). However, that region also includes ABO, suggesting that the signal can as well be driven by ABO. In agreement with this, when restricting the analyses to Trans-CDS, ABO is still highly significant ($P = 1.18 \times 10^{-45}$), in contrast to OBP2B where the P-value drops dramatically ($P = 2.49 \times 10^{-13}$). A similar pattern was seen for the ABO region for several other proteins (CTRC, ICAM2, PECAM1, PODXL, SELE).

Another interesting example is a region on chromosome 19, where over 50 SNV-sets are associated with TNFRSF10C levels. While the P-values for the Trans-Flank-sets are more similar and reach the minimum of 1.53×10^{-26}, one P-value for the Trans-CDS-sets stands out ($P = 1.64x10^{-78}$). This is the association with the Trans-CDS-set for PLAUR (Fig. 6A). For this association, it is likely that coding variants in PLAUR are driving the association. Indeed, there is also a strong association to a common variant (rs4760, MAF = 0.14, $P = 6.49x10^{-90}$) for TNFRSF10C identified in the GWAS, and adjusting for this SNV resulted in all SKAT-associations disappearing. The rs4760 is a missense variant that is annotated as deleterious (SIFT) and probably damaging (PolyPhen). Therefore, it is likely that all signals in this region were solely driven by rs4760.

Trans-associations are often driven by multiple rare variants that do not overlap with the GWAS hits

In the Trans analyses, it was evident that a larger number of associations were driven by rare variants. For example, among the SKAT associations to Flank-sets, the fraction of proteins that were identified to have an association in the rare only (Model 7) analyses (Fig. 3B) were much larger (χ^2 P-value = $2.75x10^{-09}$) for the Trans-associations (62%) than the Cis-associations (22%). An observation in the same direction was also found for the CDS-sets, but the difference was not significant (36% for Trans compared to 27% for Cis; χ^2 P-value = 0.29). Also, for 30 of the 171 individual Trans loci, analyses restricted to rare variants (Model 7) showed the most significant results (Supplementary Table S4). This is also a much larger fraction than five of the 249 loci for the Cis-regulatory regions (χ^2 P-value = $4.25x10^{-08}$). In addition, 25

out of these 30 loci with the most significant results from the rare only model (Model 7) did not overlap with a rare GWAS SNV. These are therefore unlikely to be driven by one single, but rather several rare variants. Two interesting examples are: 1) NTRK2, where the SKAT test for Trans-Flank-set around ATP2B1 resulted in a significant P-value ($P = 5.91x10^{-13}$), but where no significant GWAS SNV was identified and 2) MUC-16 where multiple rare variants in the Trans-CDS-set for MSLNL are likely to drive the association.

Also, for some rare only SKAT associations that overlapped with rare GWAS SNVs, the P-value was lower for the SKAT analyses. One such case is the association that maps to a region close to ITIH4 and TMEM110 on chromosome 3. For VWC2, the most significant SNV in the GWAS is a rare variant (MAF = 0.019, $P = 3.01x10^{-24}$) in this region. However, the SKAT analyses for the rare only variants in the Trans-CDS-set for ITIH4 resulted in a considerably lower P-value ($2.72x10^{-34}$), which indicates that additional rare variants are driving the signal (Fig. 6B). Interestingly, there are also associations to MME-levels in the same region (Fig. 6C) close to ITIH4 and TMEM110. However, for MME the Trans-Flank-sets for ITIH4 and TMEM110 are the most significantly associated, and the CommonRare analyses gives a much lower P-value compared to the rare only analyses. This suggests that the association for MME-levels is most likely driven by a common variant which agrees with the fact that the most significant GWAS SNV is, indeed, a common variant (rs35004449, MAF = 0.36).

Discussion

We have performed a large-scale gene-based association study to evaluate the combined effect of common and rare genetic variants on the expression level of 414 plasma proteins that are either well established or more exploratory biomarkers for different diseases. To some extent, the gene-based results resemble those of a single-marker GWAS, since most of the associations from the gene-based analyses overlapped with GWAS signals, and vice versa. The overlap was more pronounced for the *Cis*-associations. However, for the *Trans*-associations, there was a large number of proteins that were only associated in the gene-based analyses. This clearly highlights the potential of increasing statistical power by using gene-based tests, which can be attributed to the total number of tests being reduced by performing only few tests per gene, instead of one per genetic variant. In addition, the SKAT analyses can capture effects of multiple variants, where each one of them individually might not reach genome-wide significance in the GWAS. Interestingly, for *Trans*-associations there was also a larger number of signals that appeared to be driven by rare variants, compared to the *Cis*-associations. This could indicate a higher selective pressure, keeping protein-altering variants, such as the coding and splice site variants, at a low frequency. This agrees with a much larger fraction of rare variants having high CADD-values, which is a predictor of deleteriousness.

We tested a number of different settings in the SKAT analyses, with regards to weighting, filtering and selecting the SNVs and indels to be included in the SNV-sets. It was clear that different settings are optimal for different genetic architectures. Unsurprisingly, for regions that appeared to be driven by one single SNV or indel, the single-marker test in the GWAS appeared to be the most powerful method for

identifying the effect, which was indicated by a lower P-value. This could also be illustrated by a large number of SKAT associations disappearing when adjusting for the significant GWAS variants. However, overall, the SKAT analyses were more powerful than the GWAS since it identified additional signals that did not reach genome-wide significance in the GWAS. Associations that were driven by common variants performed better, i.e., a larger number of associations were identified, when no MAF weighting was applied in the SKAT analyses. However, signals driven by rare variants performed better using a MAF cut-off or by upweighting rare variants. Since we do not have prior knowledge of the relative contribution of rare and common variants in different regions, it is important to perform different tests to optimize the power to identify gene-phenotype associations. It is important, however, to bear in mind that increasing the number of tests requires the significance threshold to be adjusted for multiple testing by imposing stricter P-value cut-offs. Interestingly, the CADD/Eigen weighting did not dramatically influence the results compared to no weighting at all. This could suggest that the values provided by CADD and Eigen are not accurate enough for most variants to be useful in the weighting method. However, it could as well be that the differentiation between harmful and benign variants is too small to render useful as weights.

A larger number of significant associations were identified when either no MAF cut-offs or no MAF upweighting was applied. There are some possible explanations for this. First, it is possible that we are underpowered to detect the effects by rare variants, and that our results therefore are skewed towards common variants. However, it is also possible that common variants indeed account for most of the phenotypic variation seen in a population, i.e., the heritability. The latter is supported by that our results indeed showed that the total burden of damaging alleles, as predicted by CADD or Eigen values, is mostly due to common variants. This is an important observation that needs to be considered carefully in studies aiming to identify effects by rare variants.

However, it is possible that the gene-based analyses are not powerful enough to capture the effect of rare variants due to the low number of observations of rare alleles per gene. This highlights the need for even larger sample sizes in WGS studies in the future, especially for investigating the effect of rare variants in relation to complex diseases. Furthermore, the interpretation of the results would benefit from large scale simulation-based evaluations of different models and methods. However, even though the fraction of phenotypic variation explained by rare variants at population level appears to be much lower compared to common variants, rare variants can still be of high importance at the individual level. We showed a significant enrichment of rare variants among the variants with the largest impact on protein function or regulation, as predicted by CADD or Eigen scores, and also a slightly higher effect size for variants with low MAF, which supports that some rare variants have larger phenotypic effects at individual level. In a population, some individuals can have several rare alleles with high impact on a specific phenotype, which is of importance to consider when estimating, for example, polygenic risk scores.

We found that many of the SKAT associations disappeared when adjusting for the most significant GWAS hit. This could either be due to the fact that the common GWAS SNVs are indeed the causal variants, or in complete LD with them. However, there is also a possibility that the common variants capture effects of rare variants, due to LD. Multiple rare alleles that co-segregate with a more common

allele could, for example, appear to be driven by the more common allele[20]. Such interactions cannot be resolved, neither by SKAT, nor by a GWAS strategy, and it is therefore not possible to determine the true underlying architecture of the associations. SNP heritability, which is the total amount of phenotypic variation that can be captured by SNPs in a GWAS[21], has been shown to be high for many traits[2]. This suggests that GWAS SNPs might capture part of the effects caused by rare variants, even though, especially in regions with low LD, the effect of rare variants might be unmeasured due to incomplete LD with common variants analysed in the GWAS[21,22].

We have identified a number of regions that appeared to be driven by rare SNVs. There were several regions, where the SKAT models with only rare variants, or where rare variants were upweighted, seemed to be the more powerful than the other tests. One such example where SKAT identified a signal that is likely due to multiple rare variants is for AGRP. While the SKAT P-value was highly significant ($1.61x10^{-19}$), we did not identify any individual SNVs or indels in the GWAS. However, in a recent meta-analysis of GWAS summary statistics, with data from over 30,000 individuals, three independent rare SNVs were identified with the most significant SNV having a MAF of 0.0048[23]. This clearly shows that, for regions where rare variants affect the trait, MAF-weighting or filtering is indeed a powerful strategy.

Among the *Trans*-associations, several pleiotropic loci were identified. The *ABO* locus was associated with eight different proteins, but all of these were also identified in the GWAS. Another pleiotropic locus was on chromosome 3, which was associated with both VWC2 and MME. Many neighbouring SNV-sets are associated with MME levels, which agrees with previous GWAS studies[19]. Clearly, the most significant SNV-sets for MME are *ITIH4* and *TMEM110*, whereas for VWC2, only *ITIH4* stands out. Although the MME association appears to be driven by common variants, the VWC2 association is the strongest one when only rare variants are included in the SKAT test which suggests that it is not the same underlying genetic variants that drive the two associations. The most significant GWAS SNV for MME was also a common variant (rs35004449, MAF = 0.36) in complete LD with a missense variant rs4687657). However, for VWC2 a rare missense variant (rs139719930, MAF = 0.018) was the most significant. This clearly illustrates that pleiotropic effects at a locus can be due to different genetic effects that act on different pathways.

The SKAT analyses did not only provide a larger number of association signals, but there were also several regions where they provided leads towards new candidate genes. For example, in the GWAS, five proteins (BMP-6, NRP2, CD38, TNFSF13, and ADAM-TS-15) were associated with common SNVs in the same region on chromosome 5. The lead-SNVs are rs1801020 and rs2545801, which are in high LD (R^2 = 0.98), and it can therefore be assumed that their associations are driven by the same underlying genetic effects. The lead-SNVs maps to *GRK6* (G protein-coupled receptor kinase 6). However, in the SKAT analyses, both NRP2 and ADAM-TS-15 are associated with the *Trans*-CDS-set for *F12*, which is located downstream of *GRK6. F12* encodes a coagulation factor which makes it the most likely candidate in the region, and here, it is clear that the SKAT analyses perform better than GWAS to highlight candidate genes. This demonstrates the potential to identify candidate genes, at least for some proteins, by

considering only variants in the CDSs, especially when analysing gene-dense regions. However, for *Trans*-associations that are driven by non-coding variants, the ability to identify these associations would drop dramatically if analyses are restricted to CDSs only. However, most of these regions included multiple genes, and the underlying causal gene was therefore not easy to identify.

In summary, we have performed one of the most comprehensive association studies to date, aiming to identify the effect of rare and common genetic variants using WGS data, and we were able to compare several different strategies for gene-based tests. We could also clearly show that gene-based tests perform better, especially in regions where multiple rare variants contribute to the effects. However, since we do not know the genetic architecture that contributes to phenotypic variation when designing a study, it is not possible to select one test that will be the best for all regions. Therefore, it is worth highlighting that the CommonRare function in SKAT outperformed the other methods by the number of associations identified, and also performed reasonably well for associations that are driven by multiple rare variants. In conclusion, we have shown that, gene-based tests, similarly to GWAS, identify more associations to common than to rare variants. This is partly explained by the fact that a much larger fraction of the phenotypic variance explained is indeed due to common rather than rare variants. However, the power to capture effects of rare variants is limited by the low number of observations. Therefore, in future studies, when more complex phenotypes like e.g., common diseases are analysed,, substantially larger sample sizes are needed in order to identify the effect of rare variants. However, for genotype-based precision medicine interventions, where polygenic risk scores are expected to play a more central role in the future, it is of high importance to further investigate the effect of rare variants that can have a large effect on one's individual risk of developing a disease.

Materials And Methods

The Northern Sweden Population Health Study(NSPHS, N = 1069) was a health survey of the population in the Parishes of Karesuando and Soppero, County of Norrbotten, Sweden[24]. Samples (EDTA plasma and serum) were taken and immediately frozen. WGS has been performed at SciLifeLab in Stockholm, using Illumina short read technology (X-ten) to at least 30x per individual coverage, for the whole cohort following the same pipeline as described previously[5]. NSPHS was approved by the local ethics committee at the Uppsala University (Regionala Etikprövningsnämnden, Uppsala Dnr 2005:325).

Quality control of WGS data

In total, 1041 samples (Supplementary Figure S1) were sequenced and 20 individuals were removed during variant calling and quality control, as has been described previously[6], leaving 1,021 samples for downstream processing. SNV QC was performed using vcftools version 0.1.13[25]. Only biallelic SNVs and indels that did not deviate from Hardy-Weinberg equilibrium ($P > 3.85x10^{-09}$ and $P > 4.42x10^{-08}$ for SNVs and indels, respectively) were included. We also filtered on genotype quality (GQ) > 50, and finally SNVs and indels with ≥ 10% missing genotypes were removed. In addition, we removed variants in low-complexity regions based on regions identified by[26].

Protein expression levels

The protein levels for 460 putative biomarkers had been measured using the Olink Proseek Multiplex panels (CVD II, CVD III, INF I, ONC II and NEU I, www.olink.com), and the protein extension assay (PEA), as described previously[19]. Briefly, it is an affinity-based assay, where a pair of oligonucleotide-labelled antibody probes bind to the targeted protein. If the two probes are in close proximity, a PCR target sequence is formed, the resulting sequence is detected and then quantified using standard real-time PCR. The samples were analysed on ten different plates with 96 wells each. Of the 96 wells, 92 are samples, one is a negative control and three are positive controls, both used to determine the lower detection limit and to normalize the measurements. Measurements below the detection limit are removed from further analysis. In total, 903 samples were analysed, of which 892 passed the protein QC and 872 passed both protein and WGS QC and were included in the downstream analyses (Supplementary Figure S1). Protein measurements were adjusted for their position on the plate and standardized using a conservative method where the protein levels for each protein were rank-transformed to be normally distributed (mean = 0, and standard deviation = 1) within each plate. In order to achieve enough power for downstream analyses, only proteins that were above the detection level in at least 400 (46%) of the samples with WGS data were included. Also, two proteins (IL-6 and SCF) had been analysed on two different panels (ONC II and CVD II) and here the ONC II values were removed from the analyses due to lower number of individuals passing QC. After QC, 414 unique proteins remained (Supplementary Table S1) and were analysed in this study.

Annotation of genetic variants depending on their predicted deleteriousness: CADD and Eigen values

Variants were annotated using the Combined Annotation Dependent Depletion (CADD version 1.3, downloaded on 2018-05-14) database to identify what effect the variants are predicted to have on the function of the gene-product. From the CADD database, we used the PHRED-scaled CADD scores, meaning that a CADD score above 10 corresponds to the 1% most damaging variants, a CADD score above 20 to the 0.1% most damaging and a CADD score above 30, to the 0.01% most damaging variants[27]. CADD values were also estimated for indels using the online tool (https://cadd.gs.washington.edu/score, v1.3, March 2018). Variants were also annotated for their predicted effects using Eigen-PC scores v1.1, which is a weighted scoring system commonly used for non-coding variants[28]. Similar to the CADD-scores, we used the PHRED-scaled values in all analyses. Individual Eigen-PC scores were available for SNVs but not for indels. However, the Eigen-PC scores are mainly based on the underlying epigenetic pattern for each region[28] and nearby SNVs have very similar Eigen-PC scores. Therefore, for SNVs and indels with no pre-computed Eigen-PC score available, we used the score of the nearest SNV with pre-computed Eigen-PC score, if such SNVs existed within 100bp.

Annotation of genetic variants to coding, regulatory or gene-flanking regions: CDS-sets, Reg-sets, and Flank-sets

SNVs and indels were first annotated based on whether they belonged to the coding sequence (CDS) of all transcripts in GENCODE (version 26). A total of 18,467 genes had at least one SNV or indel that overlapped with any CDSs or its transcripts. For each of these genes, we constructed one SNV-set (from

now on referred to as **CDS-set**), containing all the SNVs and indels that mapped to any of its CDS, or 40bp up/downstream of a CDS, to include variants important for splicing (Supplementary Figure S3). Non-coding variants were annotated in relation to potential regulatory regions of the 405 autosomal genes encoding measured proteins. For regulatory regions, transcription starting sites (TSS) and untranslated regions (UTRs) for all possible isoforms of a gene were identified using the GENCODE annotations (version 26). A promoter annotation, defined as the region 2kb upstream of each TSS, was also included. In addition, we selected regulatory regions that overlapped with CTCF binding sites, open chromatin, transcription factor binding sites, promoters, promoter flanking regions and enhancers using Ensembl regulatory annotations[29]. A total number of 9,674 partly overlapping regulatory regions were identified for 405 autosomal genes encoding the proteins (for example, a promoter annotation as the 2kb region upstream of a TSS often overlap partly with the promoter annotation from Ensembl). For each of these genes, all SNVs and indels that mapped to any of its potential regulatory regions, except 3´UTR, were combined into one regulatory SNV-set (Supplementary Figure S3) per gene (from now on referred to as **Reg-set**). SNVs and indels that were annotated to both a CDS-set and a Reg-set were excluded from the Reg-set. The Reg- and CDS-sets above included SNVs and indels that overlapped with the regulatory regions close to a gene, or with the CDSs. However, it is possible that there are regulatory effects that might fall outside these ranges. In order to capture such effects, we also created SNV-sets consisting of all SNVs and indels within 100kb up/downstream (Supplementary Figure S3) of each gene (from now on referred to as **Flank-set**). Here, the start and stop positions (to determine if a variant is within 100kb up/downstream of the gene) were selected as the min/max coordinate for any of the regulatory regions, or any CDS for respective gene.

Cis- and Trans- analyses

The underlying hypothesis, supported by our previous studies[6], is that the expression of a protein is mainly driven by *Cis*-regulatory genetic variants. In our primary analyses, we therefore analysed both the Reg- and Flank-sets for each of the 405 autosomal genes encoding any of the proteins *in Cis* (*Cis*-Reg- and *Cis*-Flank-sets). In order to reduce the number of variants tested in the *Cis*-Flank-sets, we filtered on SNVs and indels with Eigen score > 10. Here, Eigen scores were selected instead of CADD scores, since regulatory variants are more likely to have effect on the expression of a gene *in Cis*. It is also possible that coding variants in a gene, either directly influence the expression level of the encoded protein, or affect the affinity of the antibodies used to measure the protein levels. We could therefore also expect that coding variants in the genes encoding the proteins that had been measured could influence the protein measurements directly (*in Cis*). As sensitivity-analysis, we therefore also tested for associations *in Cis* between the 405 proteins encoded by autosomal genes and the respective CDS-set (*Cis*-CDS-set).

Besides *Cis*-regulatory variants, the expression of a protein can be influenced by *Trans*-regulatory effects. *Trans*-regulatory effects can be mediated through other proteins (or functional RNAs), such as transcription factors, encoded by genes on different chromosomes or located on the same chromosome as the gene encoding the protein itself. *Trans*-regulatory effects can be either due to a coding variant that influence the function of the *Trans*-regulatory protein (or functional RNA), or a regulatory variant that

influence the abundance level of the *Trans*-regulatory protein (or functional RNA). For the analyses of *Trans*-associations, we therefore analysed the CDS-sets for each of the 18,467 genes in the genome (*Trans*-CDS-sets*)*. However, Flank-sets were also analysed in *Trans* (*Trans*-Flank-set) for the same 18,467 genes. In order to reduce the number of variants in these *Trans*-Flank-sets the SNVs and indels were filtered on CADD > 10 or Eigen > 10 which should capture the 1% most damaging variants with regards to either expression or protein function.

Trans-associations were defined as signals that did not overlap with the region surrounding the gene encoding each protein. To exclude all effects of *Cis*-SNVs (due to LD), we required *Trans*-regulatory SNV-sets to be located on a different chromosome as the gene encoding the proteins. Signals located more than 10 Mbp away from the gene on the same chromosome, unless there was a strong *Cis*-association that appeared to extend over more than 10 Mbp, were also considered being *Trans*-associations.

Statistical analyses

Analyses were performed using SKAT in R version 3.5.0. All models were adjusted for sex and age. We did choose to include a restricted number of covariates in the models, which is common in GWAS. In previous studies we have shown that several precision variables have strong effects on the levels of some of the proteins investigated[18]. However, it is unlikely that any such variables have an effect on the genetic variants and are therefore not considered being potential confounders in our study. All tests (except for the CommonRare function, for which the methodology is not implemented in SKAT) were also adjusted for relatedness by including a pairwise kinship matrix. The kinship matrix was constructed using 300,000 SNPs with MAF > 5% selected to represent tagSNPs[6]. The *Trans*-CDS and *Trans*-Flank-sets were analysed in relation to all plasma proteins, which resulted in a total of 18,467 genes times 414 proteins analysed. The *Cis*-CDS, *Cis*-Reg, and *Cis*-Flank-sets were only analysed for autosomal chromosomes and only in relation to the proteins it encoded for, which resulted in 405 gene-protein pairs analysed (Table 1).

All SNV-sets were analysed using the same seven models (Table 1). Our primary SKAT analyses included all variants that had passed QC, independent of MAF, and variants were either unweighted (model 1), weighted by their CADD/Eigen scores (model 2) or by MAF (model 3–5) with three different β-distributions: β(1, 25) where rare variants are dramatically upweighted[30] compared to common ones that are assigned almost zero weights, β(1, 5) and β(0.5, 0.5) where rare variants are slightly upweighted (Supplementary Figure S4). We then used the SKAT CommonRare function (model 6) that first analyses common and rare variants separately and then combines the test statistics[15]. We used MAF threshold equal to $1/\sqrt{(2 * \text{sample size})}$ as explained above and in[15]. The sample size differed somewhat for the proteins (Supplementary Table S1), but for the ones that had passed QC in all 872 participants this corresponds to MAF = 0.024. In the CommonRare analyses, we used the default weights, β(1,25) for rare and β(0.5,0.5) for common, where rare variants are weighted against each other with a much higher weight for very rare variants but where the common variants have a much smaller difference in weights between different allele frequencies (Supplementary Figure S4). Finally, we used a model (model 7) where only the rare variants were considered, using the CommonRare function with rareonly set to TRUE

function. We estimated the family-wise error rate for the different analyses by resampling (1000 permutations) to be between 0.0491 and 0.0504 for the CommonRare analyses and somewhat lower (0.0444–0.0465) for the other SKAT analyses. For the *Cis*-analyses, a Bonferroni-corrected P-value < $5.88x10^{-06}$ (0.05 / 405 proteins / 3 *Cis*-SNV-sets / 7 SKAT models) was used as threshold for significance (Table 1). For the *Trans*-analyses, the threshold for significance was: P < $4.67x10^{-10}$ (0.05 / 414 proteins / 2 * 18,467 *Trans*-SNV-sets / 7 SKAT models).

In order to compare the results from our SNV-sets to single-marker association results, and to be able to condition on SNVs identified in a single-marker test, we also performed a GWAS for each protein. Here, we used the same QC as for SNVs in the primary analyses and the analyses were performed using GEMMA[30] with the same covariates (age, sex) and adjustment for relatedness as in the SKAT analyses. In order to capture as many GWAS signals as possible for the conditional analyses, a liberal P-value of $5x10^{-8}$ (which is the standard threshold in a GWAS with genotyped/imputed variants) was used as the threshold for significance. For proteins with any single SNV with a P-value below that threshold, conditional GWAS were performed in order to identify additional independent single-SNV signals. This procedure was repeated until no additional significant SNV was identified. In the comparisons, with regards to the number of proteins with a significant association between SKAT and the GWAS, a more stringent threshold for multiple testing was considered. In our previous study, we have estimated the appropriate threshold for significance to be P = 0.05 / 3,078,707 independent SNVs = $1.62x10^{-8}$ for the WGS data and one phenotype[6], and the corresponding P-value for analysing 414 proteins would therefore be: $1.62x10^{-8}$ / 414 = $3.92x10^{-11}$ to reach a multiple-testing adjustment which is as strict as in our SKAT analyses. However, for *Cis*-associations, only 405 proteins and a 2Mb region up- and downstream of each of the 405 genes encoding the proteins were considered. Therefore, only 405 * 4Mb = 1,616Mb (about 53.87% of the total genome size), was analysed in total for all proteins together resulting in a P-value threshold of $3.00x10^{-8}$ to fully adjust for multiple testing.

Conditional analyses were then performed using SKAT, adjusting for common GWAS SNVs (primary and conditional SNVs). The same threshold for defining a variant as commons vs rare (MAF = $1/\sqrt{(2*\text{sample size})}$), as in the CommonRare analyses, was used. GWAS-significant SNVs with a MAF above the threshold were included as covariates in the SKAT analyses, in addition to the covariates used in the primary analyses. We performed the same set of analyses as above with the same weighting and filtering options. Conditional analyses were only performed if there was an overlap between the SKAT results and the GWAS results, i.e., if a *Cis*-regulatory SNV/SNV-set was identified with both methods, or if a *Trans*-regulatory SNV/SNV-set in the same region was identified with both methods. Also, from the SKAT analyses, only one SNV-set (with the lowest P-value) per locus was included in the conditional analyses, but all seven SKAT methods (Table 1) were used. Here, multiple-testing adjustment was applied, considering the number of proteins analysed and the seven SKAT methods.

Declarations

Author contributions. ÅJ conceived the study. MK, NR, TK and ÅJ planned and designed the study. MJ, NR, ÅJ, and DE performed analyses and MJ, NR, and ÅJ produced the figures. MC, NR, JH, HG, DS, WEE, SE, DE, BN, TK, and ÅJ wrote the manuscript.

Competing interests.

The author declares no conflict of interests.

Acknowledgements

We acknowledge all the participants and staff involved in NSPHS for their valuable contribution. The NSPHS was funded by the Foundation for Strategic Research (UG) and the European Commission FP6 (UG). Sequencing was funded by the Science for Life Laboratory (SciLifeLab), Swedish Genomes Program, which has been made available by support from the Knut and Alice Wallenberg Foundation. Sequencing was performed by NGI (National Genomics Infrastructure), Stockholm, Sweden. Protein measurements were carried out by Olink Proteomics AB in Uppsala, Sweden. The computations and data handling were enabled by resources in project SNIC 2018/8-372, sense2016007 provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was also funded by the SciLifeLab´s Technology Development Project (TDP) program, the Swedish Medical Research Council (2019-01497) and the Swedish Heart-Lung foundation (nr. 20200687). MK is financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab.

References

1. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700 000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

2. Karlsson, T. *et al.* Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat. Med.* **25**, 1390–1395 (2019).

3. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

4. Farooqi, I. S. *et al.* Clinical Spectrum of Obesity and Mutations in the Melanocortin 4 Receptor Gene. *N. Engl. J. Med.* **348**, 1085–1095 (2003).

5. Ameur, A. *et al.* SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. *Eur. J. Hum. Genet.* **25**, 1253–1260 (2017).

6. Höglund, J. *et al.* Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Sci. Rep.* **9**, 16844 (2019).

7. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res. - Fundam. Mol. Mech. Mutagen.* **615**, 28–56 (2007).

8. Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).

9. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, (2009).

10. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19**, 1563–1565 (2016).

11. Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 542 (2020).

12. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

13. Dane, F., Liu, J. & Zhang, C. *TCS: a computer program to estimate gene genealogies. Genetic resources and crop evolution* **54**, (Cambridge University Press, 2007).

14. Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *Nat. Commun.* **9**, 4674 (2018).

15. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92**, 841–853 (2013).

16. Ahsan, M. *et al.* The relative contribution of DNA methylation and genetic variants on protein biomarkers for human diseases. *PLoS Genet.* **13**, e1007005 (2017).

17. Enroth, S., Bosdotter Enroth, S., Johansson, Å. & Gyllensten, U. Effect of genetic and environmental factors on protein biomarkers for common non-communicable disease and use of personally normalized plasma protein profiles (PNPPP). *Biomarkers* **20**, 355–364 (2015).

18. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat. Commun.* **5**, 4684 (2014).

19. Enroth, S. *et al.* Systemic and specific effects of antihypertensive and lipid-lowering medication on plasma protein biomarkers for cardiovascular diseases. *Sci. Rep.* **8**, 5531 (2018).

20. Ek, W. E. *et al.* Genetic variants influencing phenotypic variance heterogeneity. *Hum. Mol. Genet.* **27**, 799–810 (2018).

21. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).

22. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).

23. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).

24. Igl, W., Johansson, A. & Gyllensten, U. The Northern Swedish Population Health Study (NSPHS)–a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* **10**, 1363 (2010).

25. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

26. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

27. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–5 (2014).

28. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).

29. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biol.* **16**, 56 (2015).

30. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–37 (2012).

Figures

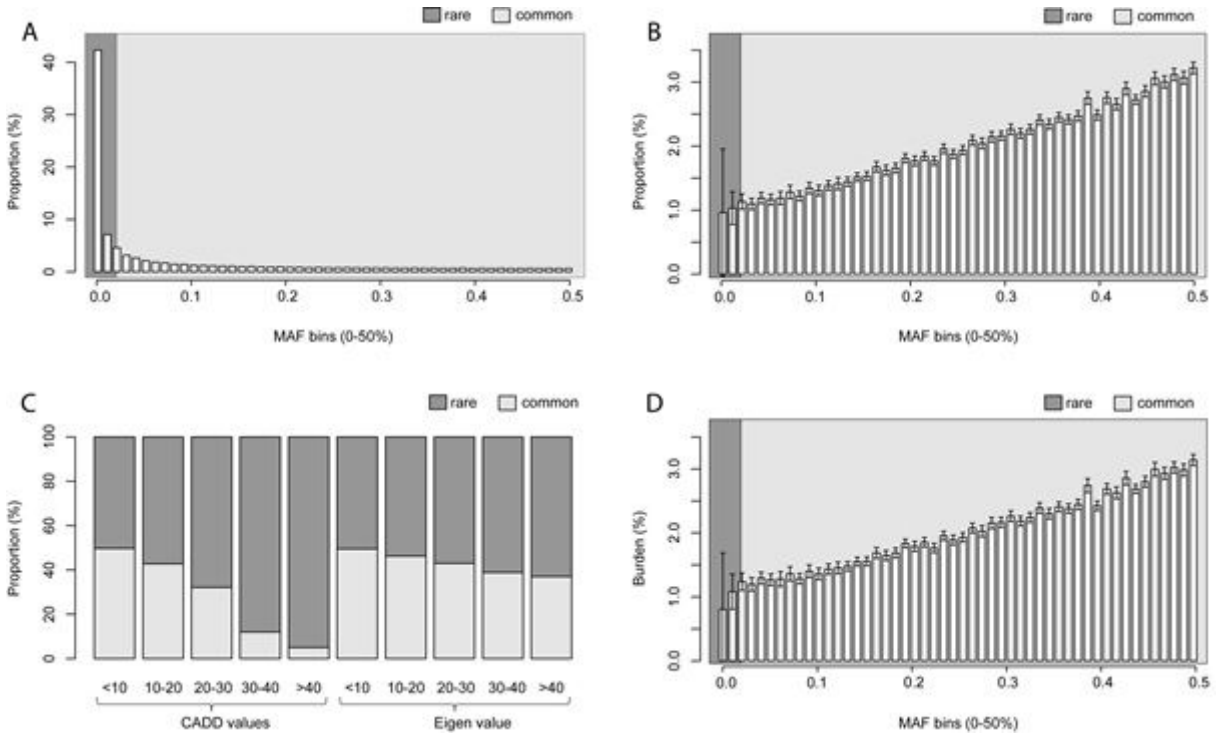

Figure 1

Distribution of MAF and CADD/Eigen values for SNVs and indels. In all figures, the dark grey area indicates the rare variants, as defined in our analyses (MAF <= 2.39%), and the light grey indicates the common variants (MAF > 2.39%). A) The MAF distribution in the cohort. The bars represent the proportion of variants with a MAF within each frequency bin. B). The MAF distribution per sample. The bars represent proportion of variants per sample belonging to different MAF bin. Averages across all individuals are shown and the error bars represent the 95% width of the distribution in the cohort. C) The fraction of the SNVs and indels being rare vs. common, for different CADD and Eigen values. D) The per-individual burden of damaging variants, as predicted by CADD in different MAF bins. Averages across all individuals are shown and the error bars represent the 95% width of the distribution in the cohort.

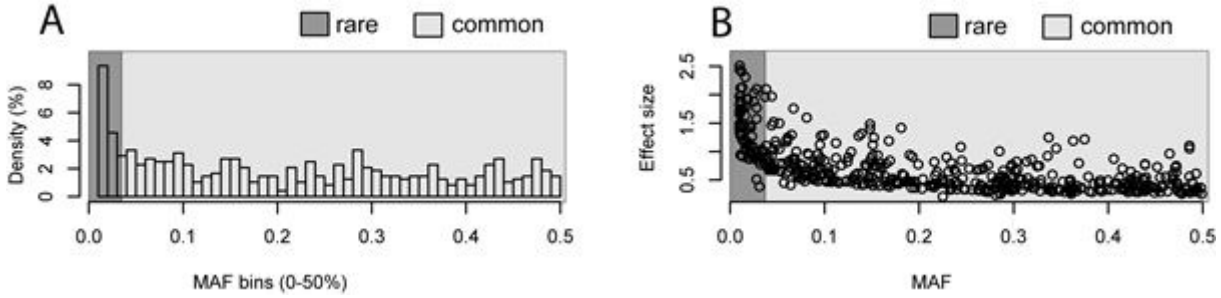

Figure 2

MAFs and effect sizes for the lead GWAS SNVs and indels. A) Distribution of MAFs for the lead GWAS significant (P<5x10-8) primary and conditional hits. A MAF-threshold of 1% was used in the GWAS, and consequently no GWAS hits had a MAF below 1%. B) Effect sizes from the GWAS, in relation to MAF for the primary and conditional GWAS hits. All effect estimates are reported as absolute values.

Figure 3

A) Overview of the SNV-sets, SKAT tests performed, and overlap between the results for the different SNV-sets. In the Venn diagram of the overlaps, for the Cis-associations, a P-value of 5.88x10-6 was considered as threshold for significance for the SKAT analyses, and for Trans-associations, a P-value of 4.67x10-10. B) Fraction of proteins with a significant hit for the different models and each SNV-sets. A total of 208, 205, 195, 50, and 66 proteins had any significant hit with the five SNV-sets respectively (N in the legend). Each bar represents the fraction of these N proteins, that were significant for the different SKAT models. The seven models are: 1) Unweighted, 2) CADD or Eigen weighted, 3) MAF weighted - β(1, 25), 4) MAF weighted – β(1, 5), 5) MAF weighted – β(0.5, 0.5), 6) CommonRare, and 7) Rare only.

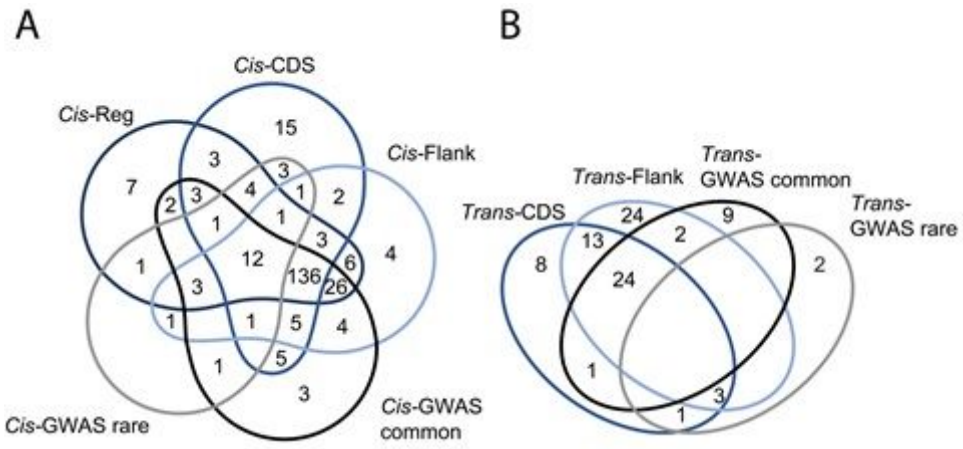

Figure 4

Overlap between proteins with a significant hit using the different SNV-sets and in the GWAS: A) Cis-associations, and B) Trans-associations. For the Cis-associations, a P-value of 5.88x10-6 was considered as threshold for significance for the SKAT analyses, and a P-value of 3.00x10-8 for the GWAS. For Trans-associations, a P-value of 4.67x10-10 was considered as threshold for significance for the SKAT analyses, and a P-value of 3.92x10-11 for the GWAS.

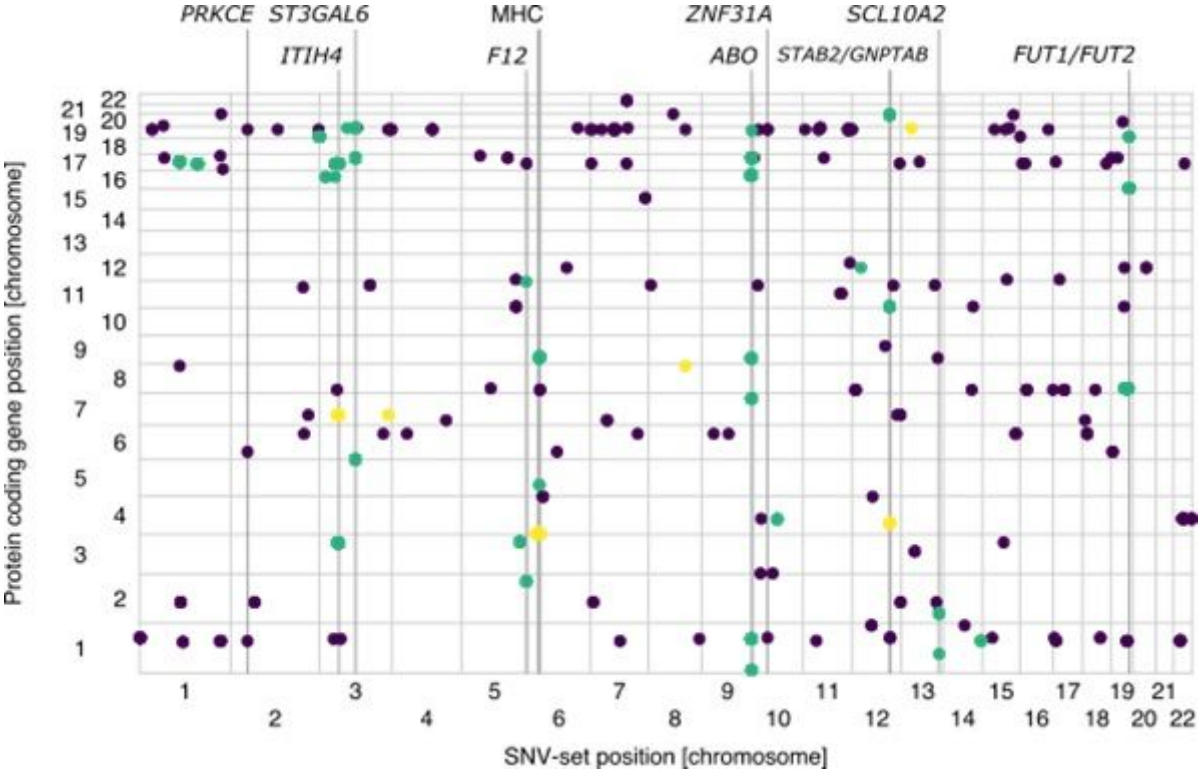

Figure 5

Locations of Trans-SKAT-signals. The x-axis shows the positions of the SNV-sets and the y-axis the positions of the genes encoding respective protein (for better separation of points, small jitter was added). SNV-sets that are associated with several proteins (pleiotropic loci) are annotated on top of the diagram. Green colour indicates overlap with a GWAS hit and yellow indicates overlap with a rare GWAS-

SNV. The MHC region, annotated as MHC on top of the diagram, contains several SNV-sets that are associated with several different proteins.

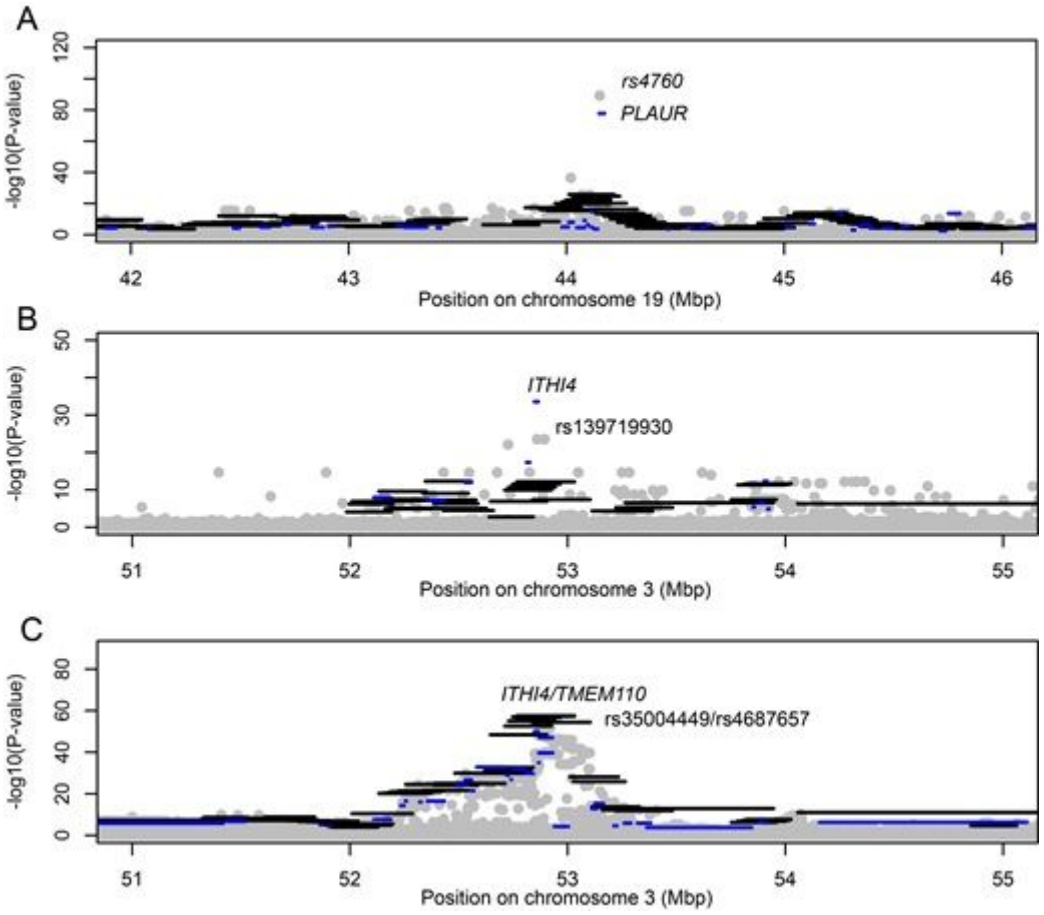

Figure 6

Regional plots for three proteins. The grey circles are the P-values from the GWAS. Horizontal lines indicate the results for the Trans-Flank-sets in black and Trans-CDS-sets in blue. As can be clearly seen, multiple, partly overlapping Trans-Flank-sets have been analysed. A) Results for TNFRSF10C levels, B) VWC2 levels, and C) MME levels

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx
- TableS3.finalreduced.xlsx
- SupplementaryFigures.pdf
- TableS4.V3reduced.xlsx