Multi-omics in research: epidemiology, methodology, and advanced data analysis

Faquih, T.O.

Chapter 6

Normal range CAG repeat size variations in the HTT gene are associated with an adverse lipoprotein profile partially mediated by body mass index

Tariq O. Faquih[1], N. Ahmad Aziz[2,3], Sarah L. Gardiner[4], Ruifang Li-Gao[1,5], Renée de Mutsert[1], Yuri Milaneschi[6,7,8,9], Stella Trompet[10], J. Wouter Jukema[11], Frits R. Rosendaal[1], Astrid van Hylckama Vlieg[1], Ko Willems van Dijk[12,13,14], Dennis O. Mook-Kanamori[1, 15]

[1] Department of Clinical Epidemiology, Leiden University Medical Center Leiden, The Netherlands; T.O.Faquih@lumc.nl (T.O.F.); R.Li@lumc.nl (R.L.-G.); R.de_Mutsert@lumc.nl (R.d.M); F.R.Rosendaal@lumc.nl (F.R.R.); A.van_Hylckama_Vlieg@lumc.nl (A.v.H.V.);
D.O.Mook@lumc.nl (D.O.M.-K.)
[2.] Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany; Ahmad.Aziz@dzne.de (N.A.A.)
[3.] Department of Neurology, Bonn University Hospital, Bonn, Germany; Ahmad.Aziz@dzne.de (N.A.A.)
[4] Department of Neurology, Amsterdam UMC Amsterdam, The Netherlands; esel_gardiner@hotmail.com (S.L.G.)
[5] Metabolon, Inc. Morrisville, North Carolina, United State of America; R.Li@lumc.nl (R.L.-G.).
[6] Department of Psychiatry, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[7] Amsterdam Public Health, Mental Health program, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[8] Amsterdam Neuroscience, Mood, Anxiety, Psychosis, Sleep & Stress program, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[9] Amsterdam Neuroscience, Complex Trait Genetics, Amsterdam, The Netherlands; y.milaneschi@amsterdamumc.nl (Y.M.)
[10] Department of Internal Medicine, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands; S.Trompet@lumc.nl (S.T.)
[11] Department of Cardiology, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands; j.w.jukema@lumc.nl (J.W.J.)
[12] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[13] Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[14] Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands; K.Willems_van_Dijk@lumc.nl (K.W.v.D)
[15] Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands; D.O.Mook@lumc.nl (D.O.M.-K.)

Manuscript submitted, under revision

1 ABSTRACT

Tandem CAG repeat sizes of 36 or more in the huntingtin gene (*HTT)* cause Huntington disease. Apart from neuropsychiatric complications, the disease is also accompanied by metabolic dysregulation and weight loss, which contribute to a progressive functional decline. Recent studies also reported an association between repeats below the pathogenic threshold (<36) for Huntington's disease and body mass index (BMI), suggesting that *HTT* repeat sizes in the non-pathogenic range are associated with metabolic dysregulation.

In this study we hypothesized that *HTT* repeat sizes <36 are associated with metabolite levels, possibly mediated through reduced BMI. We pooled data from three European cohorts (n=10,228) with genotyped *HTT* CAG repeat size and metabolomic measurements. All 145 metabolites were measured on the same targeted platform in all studies. Multilevel mixed-effects analysis using the CAG repeat size in *HTT* identified 67 repeat size-metabolite associations. Overall, the metabolomic profile associated with larger CAG repeat sizes in *HTT* were unfavorable—similar to those of higher risk of coronary artery disease and type 2 diabetes—and included elevated levels of amino acids, fatty acids, LDL, VLDL and IDL related metabolites whilst with decreased levels of very large HDL related metabolites. Furthermore, the associations of 50 metabolites, in particular specific very large HDL related metabolites, were mediated by lower BMI. However, no mediation effect was found for 17 metabolites related to LDL and IDL.

In conclusion, our findings indicate that large non-pathogenic CAG repeat sizes in *HTT* are associated with an unfavorable metabolomic profile despite their association with a lower BMI.

2 INTRODUCTION

Huntington disease (HD) is an autosomal dominant neurodegenerative disorder caused by the expansion of a cytosine-adenine-guanine (CAG) repeat in the first exon of the huntingtin gene (*HTT)*. The age of onset of the disease is determined by the number of CAG repeats in this exon: Full penetrance occurs when the number of repeats exceeds 36-39 units (1-4), while fewer than 36 repeats are considered non-pathogenic. However, repeat sizes ranging between 27 and 35 units are categorized as intermediate and have been associated with increased germline instability (4). Symptoms of HD include progressive motor, behavioral and cognitive deterioration, resulting in increasing functional decline and death within 15-20 years after disease onset (1). Intriguingly, HD is also characterized by a range of bio-energetic defects, including insulin resistance, increased sedentary energy expenditure and weight loss, despite increased appetite and caloric intake (5, 6).

The prevalence of HD is higher in populations of Caucasian descent than in Asian and African populations (3, 7). Recent estimates of the prevalence in Europeans vary from 9.7 to 17.3 per 100,000 (3, 4). In a study among five large European population-based cohorts (n~14,000), about 6.5% of the participants were found to have an intermediate or pathogenic number of CAG repeats within the *HTT* gene (8, 9). The pathophysiology of HD is complex and remains to be fully elucidated. Current findings suggest that somatic instability of tandem repeats, as well as disruption of transcriptional regulation, immune and mitochondrial function, protein trafficking, and post-synaptic signaling are likely to be involved (10, 11). Importantly, the rate of weight loss in HD was found to increase with larger CAG repeat sizes (10). Analysis of plasma, serum, and post-mortem brain samples of HD patients have found altered metabolite levels (12-14), reduced concentrations of branched-chain amino acids (15), phosphatidylcholines (15, 16), and reduced whole body cholesterol levels (11). Interestingly, CAG repeat size within the normal and intermediate range, which are considered non-pathogenic, have been associated with depression (17) and cognitive function (18). Metabolic dysregulation in HD patients implies that the CAG repeats in the *HTT* gene may directly affect systemic metabolism. However, the metabolomic signature of the highly polymorphic CAG repeat number variations in the *HTT* gene remains unexplored.

Here, we aimed to profile the metabolomic associations of *HTT* CAG repeat size variations in the non-pathogenic range, by utilizing a targeted nuclear magnetic resonance (1H-NMR) metabolomics platform. This platform included measurement of 145 metabolites, such as amino acids and lipoprotein measurements. To this end, we pooled 1H-NMR and genotype data from three large European cohorts (n=10,275). Given the aforementioned negative association between *HTT* CAG repeat size and BMI, we also aimed to assess to what extent the association between *HTT* CAG repeat size and metabolite levels is mediated through changes in BMI. We hypothesized that longer CAG repeat sizes in the *HTT* gene are associated with an unhealthy metabolomic profile, despite lowering BMI.

3 RESULTS

3.1 Population characteristics

We pooled the individual level datasets from the Netherlands Epidemiology of Obesity (NEO) (19), the Prospective Study of Pravastatin in the Elderly at Risk (PROSPER)(20), and the Netherlands Study of Depression and Anxiety (NESDA) (21) studies (N= 10,228). Characteristics of these studies are summarized in Table 1. The mean age was higher in the PROSPER study (76 years) than NEO (56 years) and NESDA (42 years). PROSPER was the only study to include participants outside the Netherlands, namely Scotland (n=1,808) and Ireland (n=1,448)). Sex distribution was skewed in the NESDA study (65.9% women, as expected due to oversampling of depressed subjects (22)), but nearly equal in NEO and PROSPER studies. Overall, the sex distribution was nearly even in the pooled dataset (54% women). Median CAG repeat sizes in both *HTT* alleles were equal in all studies (Figure 1; Figure 2).

3.2 Associations between *HTT* CAG repeat size variations and metabolite levels

Results from the multilevel mixed-effects linear regression analysis using the metabolite concentrations as the outcomes and *HTT* CAG repeat size, specifically of the longer allele, as exposure variable are presented in Figure 3 and Supplementary Table 1. *HTT* CAG repeat size in the long allele in the combined cohort were statistically significantly associated with the levels of 67/145 metabolites. These included concentrations of different branched and aromatic amino acids, fatty acids, ketone bodies, cholesterols, glycerides, phospholipids, as well as measurements related to different lipoprotein subfractions.

Overall, larger CAG repeat sizes in the long *HTT* allele were associated with increased concentrations of 59/67 metabolites. Conversely, the levels of 8/67 metabolites decreased with larger CAG repeat sizes in the long *HTT* allele.

Amino acids, fatty acids, and ketone bodies

Among the amino acids and branched amino acids, larger CAG repeat sizes in the long allele were associated with higher concentrations of alanine, glutamine, tyrosine, and valine levels. Those larger alleles were also associated with higher concentrations of total fatty acids (monosaturated and unsaturated), omega-3 fatty acids, and docosahexaenoic acid. In contrast, they were associated with lower concentrations of acetate and beta hydroxybutyrate.

3.2.1 Plasma total lipid levels

Larger CAG repeat sizes in the longer *HTT* allele were associated with increased overall serum total cholesterol concentrations — including esterified, remnant and free cholesterols. In line, larger repeat sizes were associated with increased apolipoprotein B (apoB), the apolipoprotein component found in LDL and VLDL. Moreover, measurements of phosphatidylcholine, total cholines, phosphoglycerides, and sphingomyelins concentrations increased by the longer CAG size. The larger CAG repeat sizes were not associated with serum total triglyceride levels.

3.2.2 VLDL-sized lipoproteins

Larger CAG repeat sizes in the longer *HTT* allele were also were also associated with increased total lipids of three VLDL subfractions. Specifically, larger repeat sizes were associated with increased levels of cholesterols (total, esters, and free cholesterols), total lipids, and phospholipids in very small VLDL, while levels of cholesterol esters increased with larger CAG repeat size

Figure 1

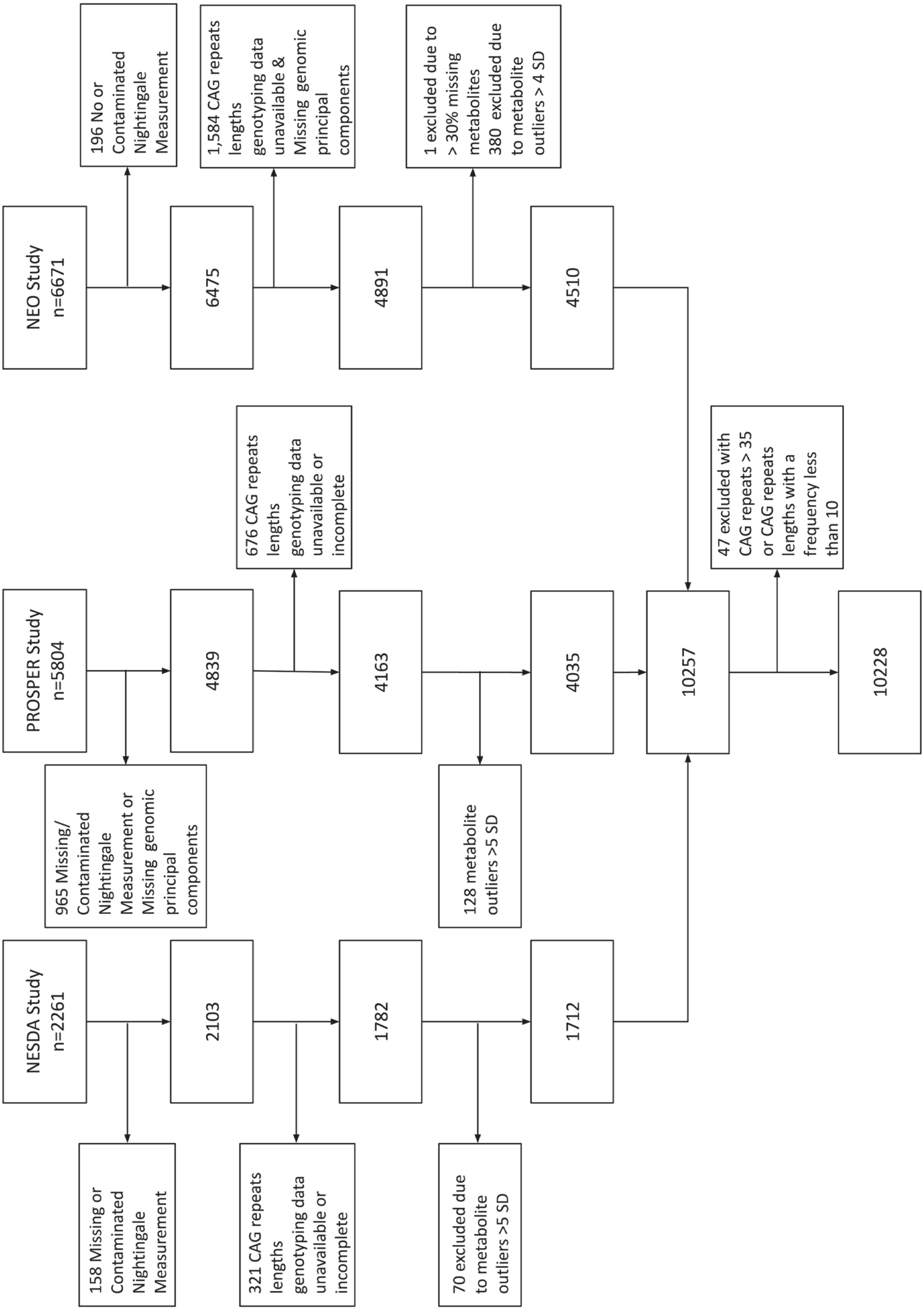

NEO Study n=6671
196 No or Contaminated Nightingale Measurement
6475
1,584 CAG repeats lengths genotyping data unavailable & Missing genomic principal components
4891
1 excluded due to > 30% missing metabolites 380 excluded due to metabolite outliers > 4 SD
4510
PROSPER Study n=5804
965 Missing/ Contaminated Nightingale Measurement or Missing genomic principal components
4839
676 CAG repeats lengths genotyping data unavailable or incomplete
4163
128 metabolite outliers >5 SD
4035
NESDA Study n=2261
158 Missing or Contaminated Nightingale Measurement
2103
321 CAG repeats lengths genotyping data unavailable or incomplete
1782
70 excluded due to metabolite outliers >5 SD
1712
10257
47 excluded with CAG repeats > 35 or CAG repeats lengths with a frequency less than 10
10228

Figure 2

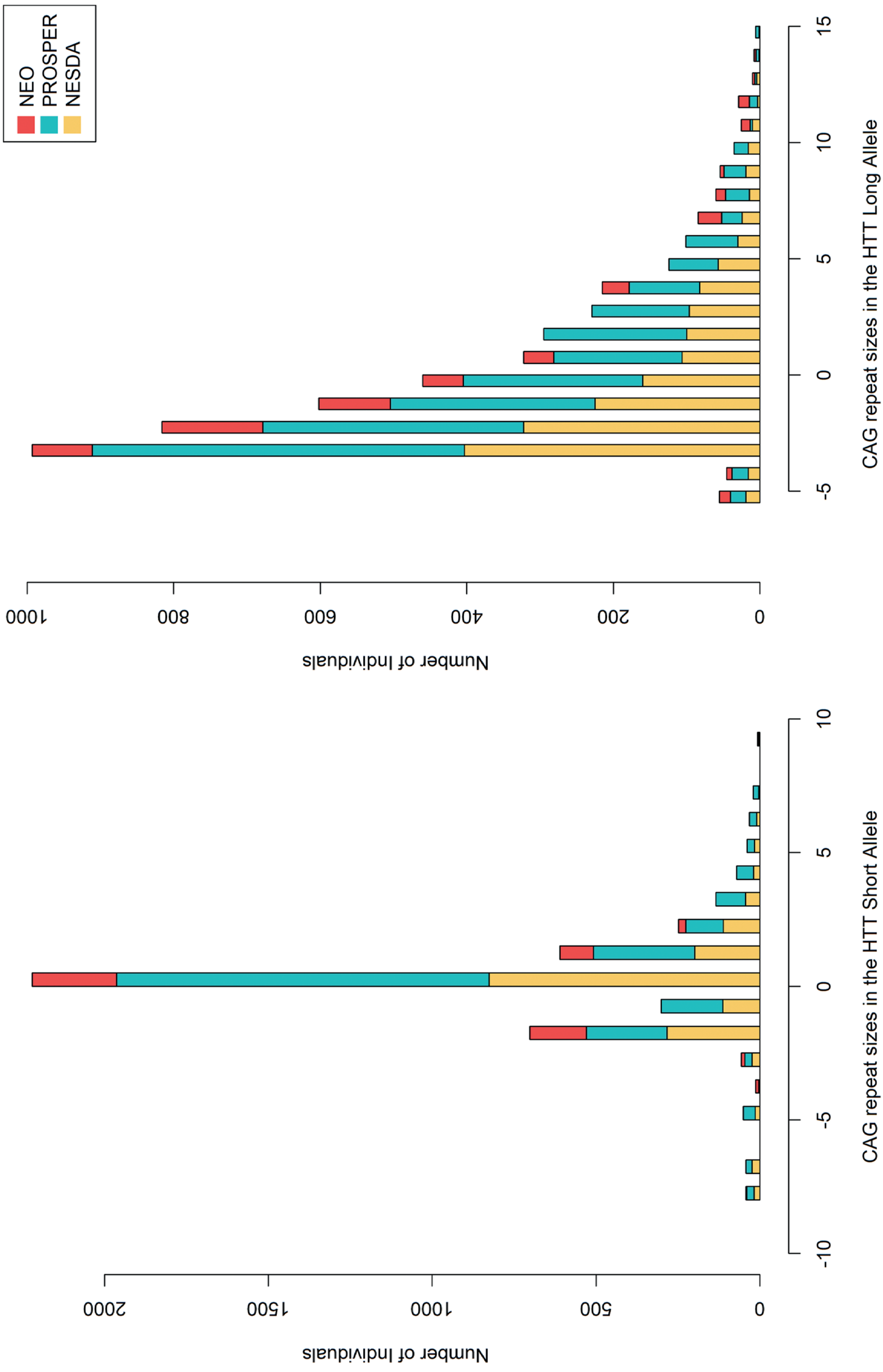
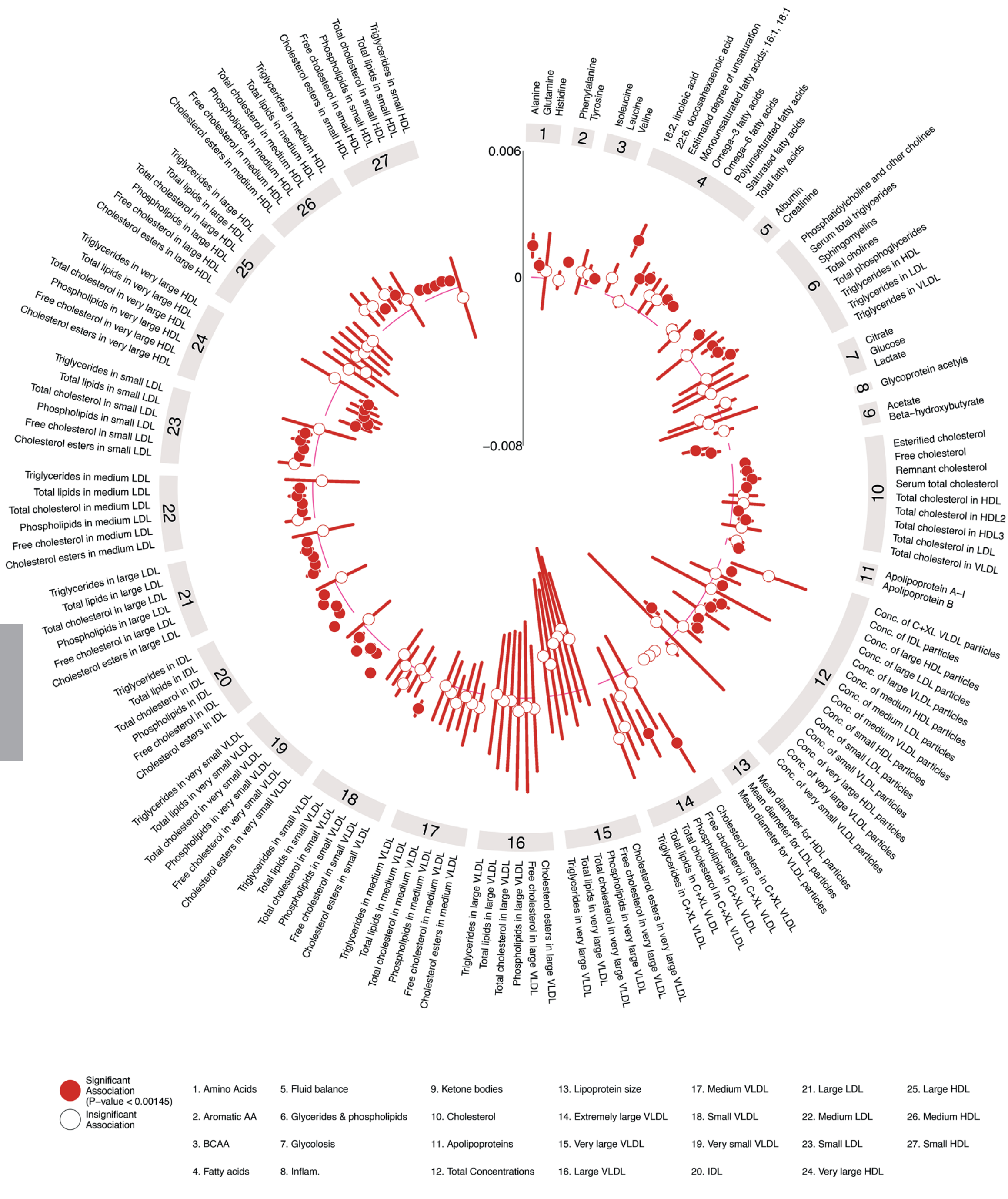

🟥 NEO
🟦 PROSPER
🟨 NESDA
Number of Individuals
0
200
400
600
800
1000
-5
0
5
10
15
CAG repeat sizes in the HTT Long Allele
Number of Individuals
0
500
1000
1500
2000
-10
-5
0
5
10
CAG repeat sizes in the HTT Short Allele

Figure 3


0.006
0
−0.008
1
Alanine
Glutamine
Histidine
2
Phenylalanine
Tyrosine
3
Isoleucine
Leucine
Valine
4
18:2, linoleic acid
22:6, docosahexaenoic acid
Estimated degree of unsaturation
Monounsaturated fatty acids; 16:1, 18:1
Omega−3 fatty acids
Omega−6 fatty acids
Polyunsaturated fatty acids
Saturated fatty acids
Total fatty acids
5
Albumin
Creatinine
6
Phosphatidylcholine and other cholines
Serum total triglycerides
Sphingomyelins
Total cholines
Total phosphoglycerides
Triglycerides in HDL
Triglycerides in LDL
Triglycerides in VLDL
7
Citrate
Glucose
Lactate
8
Glycoprotein acetyls
9
Acetate
Beta−hydroxybutyrate
10
Esterified cholesterol
Free cholesterol
Remnant cholesterol
Serum total cholesterol
Total cholesterol in HDL
Total cholesterol in HDL2
Total cholesterol in HDL3
Total cholesterol in LDL
Total cholesterol in VLDL
11
Apolipoprotein A−I
Apolipoprotein B
12
Conc. of C+XL VLDL particles
Conc. of IDL particles
Conc. of large HDL particles
Conc. of large LDL particles
Conc. of large VLDL particles
Conc. of medium HDL particles
Conc. of medium LDL particles
Conc. of medium VLDL particles
Conc. of small HDL particles
Conc. of small LDL particles
Conc. of small VLDL particles
Conc. of very large HDL particles
Conc. of very large VLDL particles
Conc. of very small VLDL particles
13
Mean diameter for HDL particles
Mean diameter for LDL particles
Mean diameter for VLDL particles
14
Cholesterol esters in C+XL VLDL
Free cholesterol in C+XL VLDL
Phospholipids in C+XL VLDL
Total cholesterol in C+XL VLDL
Total lipids in C+XL VLDL
Triglycerides in C+XL VLDL
15
Cholesterol esters in very large VLDL
Free cholesterol in very large VLDL
Phospholipids in very large VLDL
Total cholesterol in very large VLDL
Total lipids in very large VLDL
Triglycerides in very large VLDL
16
Cholesterol esters in large VLDL
Free cholesterol in large VLDL
Phospholipids in large VLDL
Total cholesterol in large VLDL
Total lipids in large VLDL
Triglycerides in large VLDL
17
Cholesterol esters in medium VLDL
Free cholesterol in medium VLDL
Phospholipids in medium VLDL
Total cholesterol in medium VLDL
Total lipids in medium VLDL
Triglycerides in medium VLDL
18
Cholesterol esters in small VLDL
Free cholesterol in small VLDL
Phospholipids in small VLDL
Total cholesterol in small VLDL
Total lipids in small VLDL
Triglycerides in small VLDL
19
Cholesterol esters in very small VLDL
Free cholesterol in very small VLDL
Phospholipids in very small VLDL
Total cholesterol in very small VLDL
Total lipids in very small VLDL
Triglycerides in very small VLDL
20
Cholesterol esters in IDL
Free cholesterol in IDL
Phospholipids in IDL
Total cholesterol in IDL
Total lipids in IDL
Triglycerides in IDL
21
Cholesterol esters in large LDL
Free cholesterol in large LDL
Phospholipids in large LDL
Total cholesterol in large LDL
Total lipids in large LDL
Triglycerides in large LDL
22
Cholesterol esters in medium LDL
Free cholesterol in medium LDL
Phospholipids in medium LDL
Total cholesterol in medium LDL
Total lipids in medium LDL
Triglycerides in medium LDL
23
Cholesterol esters in small LDL
Free cholesterol in small LDL
Phospholipids in small LDL
Total cholesterol in small LDL
Total lipids in small LDL
Triglycerides in small LDL
24
Cholesterol esters in very large HDL
Free cholesterol in very large HDL
Phospholipids in very large HDL
Total cholesterol in very large HDL
Total lipids in very large HDL
Triglycerides in very large HDL
25
Cholesterol esters in large HDL
Free cholesterol in large HDL
Phospholipids in large HDL
Total cholesterol in large HDL
Total lipids in large HDL
Triglycerides in large HDL
26
Cholesterol esters in medium HDL
Free cholesterol in medium HDL
Phospholipids in medium HDL
Total cholesterol in medium HDL
Total lipids in medium HDL
Triglycerides in medium HDL
27
Cholesterol esters in small HDL
Free cholesterol in small HDL
Phospholipids in small HDL
Total cholesterol in small HDL
Total lipids in small HDL
Triglycerides in small HDL
● Significant Association (P−value < 0.00145)
○ Insignificant Association
1. Amino Acids
2. Aromatic AA
3. BCAA
4. Fatty acids
5. Fluid balance
6. Glycerides & phospholipids
7. Glycolysis
8. Inflam.
9. Ketone bodies
10. Cholesterol
11. Apolipoproteins
12. Total Concentrations
13. Lipoprotein size
14. Extremely large VLDL
15. Very large VLDL
16. Large VLDL
17. Medium VLDL
18. Small VLDL
19. Very small VLDL
20. IDL
21. Large LDL
22. Medium LDL
23. Small LDL
24. Very large HDL
25. Large HDL
26. Medium HDL
27. Small HDL

in small VLDL. Finally, the larger CAG repeat size was also associated with increased levels of cholesterol esters and phospholipids in extremely large (XL) VLDL.

IDL-sized lipoproteins

Total concentrations of IDL increased with larger CAG repeat size in the long *HTT* allele. Likewise, the levels of total lipids, cholesterol (total, ester, and free cholesterols), and phospholipids also increased with larger CAG repeat size.

3.2.3 LDL-sized lipoproteins

Larger CAG repeat sizes in the long allele were also associated with higher concentrations of LDL-cholesterol. This association was reflected in increased levels of cholesterols (total and free) in medium and small LDL, and cholesterols (total, ester, and free) in large LDL. Furthermore, larger *HTT* CAG repeat size was associated with increased levels of total lipids and phospholipids in all three subfractions of LDL.

3.2.4 HDL-sized lipoproteins

Larger *HTT* CAG repeat sizes were associated with increased levels of total cholesterol in HDL3, which was reflected by increased levels of small and medium HDL. In small HDL, larger CAG repeat size was associated with increased levels of cholesterol (total, ester, and free), total lipids, and phospholipids. In medium HDL, the levels total lipids and phospholipids were also increased. In contrast, larger *HTT* CAG repeat sizes were related to decreased levels of very large HDL. In addition, larger *HTT* CAG repeat sizes were associated with decreased levels of all metabolites—cholesterol (total, ester, and free), total lipids, and phospholipids—in very large HDL. No associations were present between *HTT* CAG repeat sizes and apolipoprotein A-I levels, a major component of HDL particles.

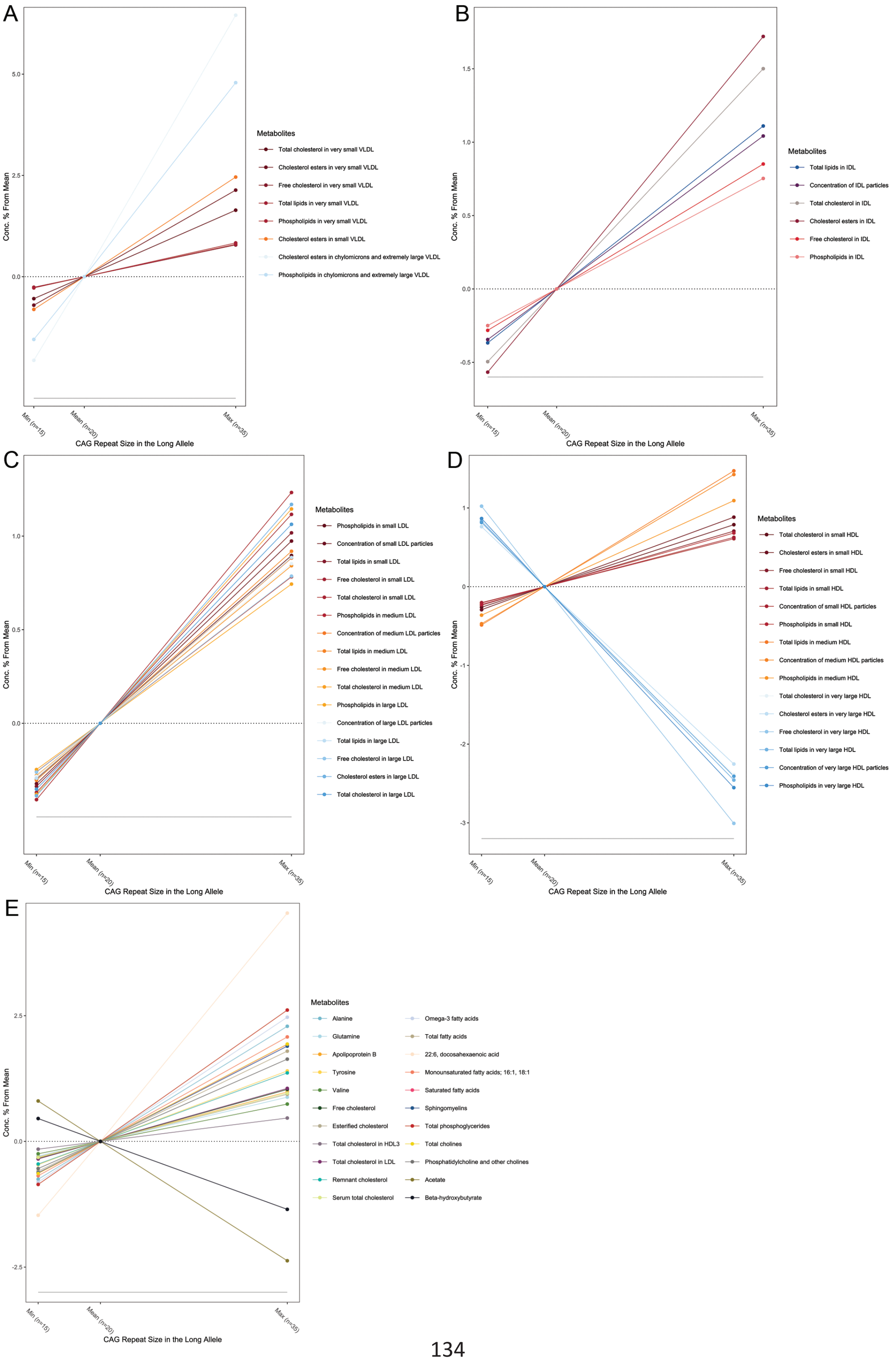
3.3 Estimation of Metabolite levels at the Largest and Smallest CAG Repeat Size

Results for the estimated percentage change for the 67 metabolites previously found in multilevel mixed-effects linear regression analysis are provided in Supplementary Table 4. Overall, at a size of 35 CAG repeats, our model predicts an increase of between 1 to 6% in the levels of all VLDL metabolites (Figure 4A). Levels of phospholipids and cholesterol esters in the XL VLDL were particularly increased to up to 5% and 6% from the mean, respectively. Levels of metabolites related to IDL, LDL, and small and medium HDL increased by 1% from the mean CAG size (Figure 4B-D). Conversely, levels of very large HDL and its lipid and cholesterol all decreased by approximately 2-3% at 35 CAG repeats compared to the mean size (Figure 4D). Amino acids, fatty acids, total cholesterols, and the other remaining metabolites, were increased between 1% and 4.5% at 35 CAG repeat. The exceptions were acetate and beta-hydroxybutyrate, which decreased by 2.4% and 1.3% at 35 CAG repeat size respectively (Figure 4E).

3.4 Nonlinear Associations

Additional sensitivity analyses were performed to assess potential interaction effects between the CAG repeat sizes in the two *HTT* alleles, as well as to assess their potential nonlinear associations with metabolite levels. In this analysis we identified 77 metabolites associated with the CAG repeat sizes, quadratic terms, or the interaction term between the two allele sizes. Of these associations, 14 CAG-metabolite associations were not previously found in the multilevel mixed-effects linear regression analysis (when the interaction and quadratic terms were not included). Ten of these 14 metabolites had an association with the quadratic terms or the inter-

Figure 4


A
5.0
2.5
0.0
Conc. % From Mean
Min (n=15)
Mean (n=20)
Max (n=35)
CAG Repeat Size in the Long Allele
Metabolites
Total cholesterol in very small VLDL
Cholesterol esters in very small VLDL
Free cholesterol in very small VLDL
Total lipids in very small VLDL
Phospholipids in very small VLDL
Cholesterol esters in small VLDL
Cholesterol esters in chylomicrons and extremely large VLDL
Phospholipids in chylomicrons and extremely large VLDL
B
1.5
1.0
0.5
0.0
-0.5
Conc. % From Mean
Min (n=15)
Mean (n=20)
Max (n=35)
CAG Repeat Size in the Long Allele
Metabolites
Total lipids in IDL
Concentration of IDL particles
Total cholesterol in IDL
Cholesterol esters in IDL
Free cholesterol in IDL
Phospholipids in IDL
C
1.0
0.5
0.0
Conc. % From Mean
Min (n=15)
Mean (n=20)
Max (n=35)
CAG Repeat Size in the Long Allele
Metabolites
Phospholipids in small LDL
Concentration of small LDL particles
Total lipids in small LDL
Free cholesterol in small LDL
Total cholesterol in small LDL
Phospholipids in medium LDL
Concentration of medium LDL particles
Total lipids in medium LDL
Free cholesterol in medium LDL
Total cholesterol in medium LDL
Phospholipids in large LDL
Concentration of large LDL particles
Total lipids in large LDL
Free cholesterol in large LDL
Cholesterol esters in large LDL
Total cholesterol in large LDL
D
1
0
-1
-2
-3
Conc. % From Mean
Min (n=15)
Mean (n=20)
Max (n=35)
CAG Repeat Size in the Long Allele
Metabolites
Total cholesterol in small HDL
Cholesterol esters in small HDL
Free cholesterol in small HDL
Total lipids in small HDL
Concentration of small HDL particles
Phospholipids in small HDL
Total lipids in medium HDL
Concentration of medium HDL particles
Phospholipids in medium HDL
Total cholesterol in very large HDL
Cholesterol esters in very large HDL
Free cholesterol in very large HDL
Total lipids in very large HDL
Concentration of very large HDL particles
Phospholipids in very large HDL
E
2.5
0.0
-2.5
Conc. % From Mean
Min (n=15)
Mean (n=20)
Max (n=35)
CAG Repeat Size in the Long Allele
Metabolites
Alanine
Glutamine
Apolipoprotein B
Tyrosine
Valine
Free cholesterol
Esterified cholesterol
Total cholesterol in HDL3
Total cholesterol in LDL
Remnant cholesterol
Serum total cholesterol
Omega-3 fatty acids
Total fatty acids
22:6, docosahexaenoic acid
Monounsaturated fatty acids; 16:1, 18:1
Saturated fatty acids
Sphingomyelins
Total phosphoglycerides
Total cholines
Phosphatidylcholine and other cholines
Acetate
Beta-hydroxybutyrate

action terms, which included citrate, apolipoprotein A-I, histidine, leucine, unsaturated fatty acid levels, mean diameters of HDL and VLDL, and measurements in large HDL and very large VLDL. Full results for these metabolites are presented in supplementary Table 2.

Among the metabolites associated with CAG repeats in the linear and nonlinear analyses (63/77), 26 had a significant association with the quadratic or interaction terms. Ten metabolites were associated with both alleles and the interaction term, and 7 metabolites had an association with the quadratic terms. These nonlinear and interaction associations were primarily with HDL, glycerides and phospholipids, fatty acids, histidine, and alanine. Overall, the associations between the linear and nonlinear models were minimal and the quadratic and interaction estimates were small.

3.5 Mediation analysis

First, we performed the multilevel mixed-effects regression between the larger *HTT* CAG repeat size as the independent variable and BMI as the outcome. Accordingly, larger CAG repeat size in the long allele was associated with lower BMI (effect estimate of -0.03 kg/m^2 per CAG repeat; 95% CI: -0.05 – -0.01; SE: 0.01).

Second, as the larger CAG repeat size were negatively associated with BMI, we performed an analysis for mediator-outcome, and calculated the mediation (indirect) and the total effect. BMI had a significant mediation effect in 50/67 of the CAG-metabolite associations. Inconsistent mediation effects, in which the direct and mediated effects were in opposing directions, were present in 74% of the associations. Metabolites without mediation effect by BMI were predominately total, free, and esterified cholesterol, total lipids, and phospholipids in LDL subfractions and IDL. The mediation effect accounted for the majority of the total effect on the concentration of very large HDL and its lipid and cholesterol content (Table 2). These measurements were also found to be strongly correlated with each other and with the total cholesterol in HDL as illustrated in the heatmap in Supplementary Figure 1. Despite the positive mediation effect on the metabolite levels, the total effect of CAG repeats on these HDL metabolites remained negative in contrast, both the mediation effect and direct effect were positive and increased the level of total cholesterol in HDL3. In summary, the overall effects of larger CAG repeat size on metabolite levels were slightly reduced after accounting for the mediation effect of BMI. Detailed results for the mediation analysis results are provided in Supplementary Table 3A and 3B.

4 DISCUSSION

We present a study on the association between CAG repeat size in the *HTT* gene—in the non-pathogenic range— and metabolite levels in more than 10,000 individuals of European ancestry. Larger *HTT* CAG repeat size in the longer allele were associated with the levels of 67 out of 145 measured metabolites. We found that the association between larger *HTT* CAG repeat sizes and total concentrations of lipid species in very large HDL remained negative despite significant mediation by lower BMI. Partial mediation by BMI was also found for 50 metabolites, wherein the larger CAG repeat sizes were associated with increased levels of lipids in small HDL and VLDL, as well as elevated levels of amino acids and fatty acids, despite inconsistent mediation by BMI. Conversely, BMI did not mediate the effects of the larger *HTT* CAG repeat sizes on the levels of 17 other metabolites, primarily consisting of cholesterol and lipids in IDL and LDL.

Overall, our findings indicate a role for tandem repeat polymorphisms in the *HTT* gene in the regulation of a diverse array of metabolites. We found that the larger size of CAG repeats of the long allele was related to increased levels of small and medium HDL, and their cholesterol content, as well as increasing omega-3 fatty acids, and total cholesterol in HDL3. In this respect, the larger CAG repeat size is related to a more favorable lipoprotein profile, such as, for example, observed during weight loss (23). However, also an unfavorable metabolomic profile with increasing long allele CAG repeat size was observed: LDL, IDL, VLDL particles, apoB, remnant cholesterols, total cholesterols, valine, tyrosine, alanine, and total fatty acids were all positively associated with the larger CAG repeat size in the long allele. The associations with HDL particles of different subfractions were heterogeneous. Opposite effect directions were found for very large HDL cholesterols, lipids, and concentration in comparison to small and medium HDL. Our mediation analyses indicated an inconsistent mediation by BMI with respect to very large HDL cholesterols and lipid levels, which were highly correlated with the total levels of HDL cholesterol. On the other hand, the associations with several LDL and IDL cholesterols levels were not mediated by BMI. Moreover, the mediation effect through BMI was generally inconsistent with the direct effect and was partial or low for 50 metabolites. These findings thus suggest the existence of an alternative pathway, independent of BMI, through which *HTT* CAG repeat size variations could affect the levels of these metabolites, including LDL and IDL.

We found that larger *HTT* CAG repeat sizes were associated with a metabolic profile similar to what was recently described in people at high risk for coronary artery disease (CAD) (24). In particular an inverse association between the cholesterols in larger HDL particles—and not the small or medium HDL—with the incidence of CAD has been reported (24). Furthermore, elevated total concentrations and cholesterols levels in LDL, IDL, VLDL, triglycerides, and apoB were accompanied by a higher incidence of CAD and peripheral artery disease (PAD). ApoB in particular has recently been reported as a strong lipoprotein marker for cardiovascular risk (25). Our findings for the associations with amino acid and branched amino acids, and fatty acids – specifically for alanine, valine, tyrosine, and total fatty acids – are also indicative of a metabolic profile associated with a higher risk of CAD, type 2 diabetes (26, 27), unhealthy adiposity (23), metabolically unhealthy normal weight (28), and inactivity (29). These heterogenous metabolomic profiles are also comparable to what has been observed in HD patients (5, 10, 30, 31), in whom weight loss and increased resting estate energy expenditure are accompanied by a higher risk of CAD and type 2 diabetes.

We additionally estimated that relatively large CAG sizes can substantially decrease very large HDL-related metabolites (by up to 3%), while increasing the levels of other lipoprotein metabolites by 1-6% as compared to the mean CAG repeats size (Figure 4). Thus, our findings indicate that larger *HTT* CAG repeat sizes result in a metabolic profile reminiscent of that associated with high CAD risk, suggesting a possible role of CAG repeats in *HTT*, and potentially other genes with polymorphic CAG repeat tracts, as genetic modifiers of clinically relevant cardiometabolic traits and disorders. Indeed, *HTT* CAG tandem repeat polymorphisms may account for part of the (missing) heritability of different metabolites, and by extension, of other phenotypes, such as BMI (32) and CAD. Thus, CAG repeat size polymorphisms are promising targets for further exploration in future studies.

4.1 Strengths and Limitations

Our study has several strengths. First, the genotyping methodology used in the three cohorts was specifically designed to genotype the tandem repeat region in *HTT*. Second, we pooled the targeted metabolomics and genotyping data from 3 European cohorts for analysis, resulting in a uniquely large sample size. Third, few metabolomic studies have been conducted in HD patients. Moreover, these studies used case-control study designs with small sample sizes (33). Our study is the largest metabolomics study thus far on the metabolomic signature associated with CAG repeat size variations in the *HTT* gene. Fourth, we found largely positive CAG-metabolite associations despite the lowering of BMI in the mediation analysis. Our study also has some potential limitations. Our study populations were at an increased risk of cardiovascular diseases and depression, which may have induced collider bias. Although, for the NEO study, we accounted for oversampling of overweight individuals, this was not possible for other characteristics, such as depression, in all studies. This was due to the unknown proportion of oversampling. However, the effect estimates were similar across studies, making it unlikely that this oversampling affected the results of our analysis. *HTT* CAG repeat size variations were also associated with the odds of depression in a previous study (17). Therefore, examining the potential role of depression may provide further insights into the mechanisms underlying the CAG-metabolite associations. However, this was beyond the scope of the current study. Finally, we could not deduce causal associations from the mediation analysis due to the difficulty verifying that no mediator-outcome confounding was present.

5 CONCLUSION

In conclusion, we examined the relationship between CAG repeat size in *HTT* with the levels of a large number of circulating metabolites. We found that non-pathogenic CAG repeat size variations in *HTT* are associated with the levels of 67 metabolites, exhibiting a heterogenous metabolomic signature. Favorable associations included positive associations with levels of cholesterol in small and medium HDL. Despite the observation that larger *HTT* CAG repeat sizes were associated with lower BMI and a favorable profile for some metabolites, we observed an additional unfavorable metabolomic profile, including associations with elevated LDL and IDL cholesterols, reduced cholesterol in very large HDL and elevated amino and fatty acids. This unfavorable profile was found to overlap with the profile seen in unhealthy adiposity, CAD, and type 2 diabetes. Based on mediation analysis, 50 metabolites showed only partial mediation and 17 metabolites—related to LDL and IDL cholesterol levels— showed no significant BMI mediation at all. Our mediation results therefore imply the potential existence of a BMI-independent mechanism underlying their association with CAG repeat size. We also found intriguing novel associations of CAG repeat size in *HTT* with metabolic dysregulation, with and without the mediation of BMI. Thus, tandem repeat polymorphisms in *HTT* and other genes may contribute to the heritability of cardiometabolic diseases and be instrumental in the elucidation of their underlying metabolomic mechanisms.

6 METHODS

6.1 Study Design

Data derived from three European cohorts were merged for pooled analyses, i.e., NEO, PROSPER, and NESDA. Details regarding the inclusion criteria of each cohort are summarized in Figure 1.

6.1.1 NEO

The NEO study is an ongoing population-based, prospective cohort study of individuals aged 45–65 years, with an oversampling of individuals with overweight or obesity. Men and women aged between 45 and 65 years with a self-reported BMI of 27 kg/m^2 or higher, living in the greater area of Leiden (in the West of the Netherlands) were eligible to participate in the NEO study. In addition, all inhabitants aged between 45 and 65 years from one municipality (Leiderdorp) were invited irrespective of their BMI, allowing for a reference distribution of BMI. Recruitment of participants started in September 2008 and completed at the end of September 2012. In total, 6,671 participants have been included, of whom 5,217 with a BMI of 27 kg/m^2 or higher. Participants were invited to come to the NEO study center of the Leiden University Medical Center for a baseline study visit after an overnight fast of at least 10 hours. During the visit fasting blood samples were taken from the participants (19). The study was approved by the medical ethical committee of the Leiden University Medical Center. The sample size for this analysis was 4,510 participants of European ancestry after exclusion of participants without metabolomic data (n=99), flawed metabolomic measurements (due to high peroxide or high ethanol content) (n=97). Moreover, as the NEO study had a higher number of extreme values than the other included studies, individuals were excluded if they had metabolite measurements above 4 standard deviations (n=380), instead of the 5 standard deviations cutoff used in the PROSPER and NESDA studies. Finally, we excluded individuals without genotype data (n=1,584). In addition, one individual was excluded due to abnormally high number of missing metabolite measurements (49/145; 33%) (Figure 1).

6.1.2 PROSPER

PROSPER was a randomized, double-blind, placebo-controlled trial among 5,786 men and women between 70-82 years old with a pre-existing vascular disease or a raised risk for such a disease. Participants were recruited from three countries with 2,517 individuals from Scotland, 2,173 individuals from Ireland and 1,096 individuals from the Netherlands. Fasting blood sample were collected and stored at -80 °C for later NMR metabolomics analysis (34). The study was approved by the institutional ethics review boards of all centers and written informed consent was obtained from all participants (20). The final sample size used in this analysis was 4,035 after exclusion of participants with flawed metabolomic data (n=965), individuals with metabolite measurements above 5 standard deviations (n=128), and participants without genotype data (n=676) (Figure 1).

6.1.3 NESDA

NESDA is an ongoing longitudinal cohort study into the long-term course and consequences of depressive and anxiety disorders. The sample consists of 2,981 participants with depressive/anxiety disorders and healthy controls recruited from the general population, general practices, and secondary mental health centers (21). Blood samples were collected after an overnight fast at the baseline visit (2004-2007). For the present analyses we initially selected data from 2,261 unrelated individuals of European ancestry identified using GWAS data. The Ethical Committees of all participating universities approved the NESDA project, and all participants provided written informed consent (35). We excluded 158 individuals without metabolomic data (n=39) or with flawed samples (n=199), and with metabolite outliers above 5 standard deviations (n=70). In addition, individuals with genotype data were also excluded (n=321). The final sample size used in the analysis was 1712 (Figure 1).

6.2 Genotyping

Due to the technical limitation of next-generation short-read sequencing to accurately call DNA repeat sequences (36), a multiplex polymerase chain reaction (PCR) method was developed using TProfessional thermocycler (Biometra, Westburg) with labelled primers to genotype the CAG repeat sizes in the two *HTT* alleles. Full details about the genotyping methodology have been described previously (17).

6.3 Metabolomics Measurements

Metabolomic profiles were measured using the Nightingale (Nightingale Health Ltd, Helsinki, Finland) NMR platform in all selected participants. Nightingale uses a targeted metabolomics approach by defining the specific metabolites to be quantitively measured in advance. This approach yields consistent and reproducible concentration measurements across studies (37). The platform measures approximately 226 metabolites and metabolite ratios, consisting predominantly of very low density (VLDL), intermediate density (IDL), low density (LDL), and high density (HDL) lipoproteins. Those lipoproteins—with the exception of IDL—are further subclassified based on their lipid composition and particle sizes(38). Accordingly, VLDL is divided into very small, small, medium, large, very large, and extremely large subfractions; HDL is divided into small, medium, large, and very large subfractions; and LDL is divided to small, medium, and large subfractions. The supplementary ratio variables calculated the ratio of various metabolites concentrations within lipoprotein subfractions, e.g., "triglycerides to total lipids ratio in IDL". Additionally, the platform measured the concentrations of various individual metabolites beyond lipoproteins such as amino acids, free fatty acids, and ketone bodies (39). For our study we excluded the 81 ratio variables and focused on the remaining 145 metabolite concentrations that were available in all 3 cohorts. Samples in all cohorts were taken after a fasting period.

6.4 Statistical Analysis

6.4.1 Multilevel mixed-effects linear regression

We performed a joint polynomial multilevel mixed-effects linear regression using data from all three cohorts. First, for each individual, we defined the *HTT* allele with the larger CAG repeat tract as 'long', and the other one as 'short'. This was done as the two alleles can have independent effects as demonstrated in previous studies (32). The number of repeats in each allele were then mean-centered to reduce multicollinearity and ease the interpretation. In order to address possible heteroscedasticity we used robust standard errors for the analysis. Influential data points (i.e., influential outliers) were accounted for by removing CAG lengths with a frequncy less than 10 in the combined cohort (n=47). Therefore, the final pooled number of particpants used in the analysis was n = 10,228 (Figure 1).

Metabolite variables were natural log-transformed and the missing values were imputed using the K-nearest neighbor imputation method described in our previous work (40, 41). In brief, for each metabolite with missingness, we selected 10 correlated metabolites with no missingness. We then used these metabolites to impute the missing values by calculating the means. We expect that this imputation method will have negligible bias and error as the number of missing values was low and sample sizes were large, as was demonstrated in the simulation results in our past work as well (41).

Since we had access to the individual level data of all three studies, we were able to perform pooled analyses, rather than meta-analyzing the effects per study. For the analysis we adjusted for age, sex, and the first 4 genetic principal components as the fixed factors. In addition, we used country and study variables as random factors in the mixed effects model. As the NEO data had an oversampling of overweight individuals, we weighted the analyses to the BMI distribution of the Dutch general population. The weight was set to 1 for the PROSPER and NESDA participants. To account for population stratification, we used the country (Netherlands, Scotland, and Ireland) and the cohort (NEO, PROSPER, and NESDA) as random effect variables.

Both alleles were included in the regression models as previous studies reported differing associations between the "long" and "short" alleles in *HTT* with different outcomes (17, 18, 32, 42). However, due to the dominant effect of the *HTT* repeat expansion in HD, we focused on the "long" allele effect estimates only. We performed the analysis for each of the 145 metabolites as the outcomes and the mean-centered number of repeats in both *HTT* alleles as the independent variables. Effects of CAG repeats have been shown to have nonlinear associations and interactions have been described between the two *HTT* alleles (17, 18, 43). Therefore, we conducted a secondary analysis to check nonlinearity and used a polynomial model by adding quadratic terms for each allele and an interaction term between the two alleles (long and short). We adjusted for age, sex, and the first 4 genetic principal components as the fixed effects and used country and study as random effects in the mixed effects model.

Data preparation and analysis were conducted with R version 4.1.0 (44). Circular plots for the effect estimates were designed using the EpiViz R package (45-47). Multilevel mixed-effects model and mediation analyses were performed by utilizing the "mixed" command in STATA/SE version 16 (StataCorp LLC) (48).

6.4.2 Multiple testing correction

To adjust for multiple testing, we used the VeffLi estimate described by Ji and Li (22). This method takes the covariance between metabolite levels into account by estimating the effective number of independent variables. Accordingly, the effective number of independent variables was 35 and the adjusted p-value cut-off was put at 0.05/35 = 0.0014.

Estimation of Metabolite levels at the Largest CAG Repeat Size

The effect estimates from the multilevel mixed-effects linear regression accounted for the effect of 1 CAG repeat size increase. By using the effect estimates per CAG repeat from the mixed linear regression model, we were able to show a simple estimation of the percentage difference from the mean of metabolites that were associated with the larger CAG repeat size from the multilevel mixed-effects linear regression analysis. We estimated the percentage change in metabolite levels at CAG repeat sizes equal to the smallest, mean, and largest CAG repeat size in the pooled dataset, corresponding to 15, 20, and 35 repeats respectively. Plots for visualizing the percentage changes were generated using the *looplot* R package (49, 50).

6.4.3 Mediation Analysis

To test for mediation by BMI of the CAG-metabolite associations, we performed three analyses as proposed by Baron & Kenny (1986) (51). First, we modelled the exposure-mediator relationship by using the multilevel mixed-effects linear regression to assess the association between the CAG repeat sizes and BMI. Second, we calculated the mediation effect using the multilevel mixed-effects linear regression for the metabolites that were associated with *HTT* CAG repeat size in the previous analysis. The natural logarithm of the metabolite levels was used as the outcome and the independent variables were the CAG repeat sizes in the short and long alleles, as well as BMI, the mediator. Third, given our large sample size, we used the simpler Sobel's test (Equation 1) instead of bootstrapping to test the mediation effect of BMI (51-53).

$$A \times B / \sqrt{}(A^2 \times B_{se}^2) + (B^2 \times A_{se}^2) + (A_{se}^2 \times B_{se}^2)$$

Equation 1: The Sobel's equation for testing mediation. A: the estimate between CAG repeat sizes in *HTT* and BMI; B: the estimate between BMI and metabolite levels; A_{se}: standard error of A; B_{se}: standard error for B; A × B is the indirect effect of BMI

Using this method, we calculated the indirect effect through BMI by multiplying the estimates of BMI from the exposure-mediator model and mediator-outcome model. We also calculated the total effect for the model by adding the direct effect, i.e. estimates of the CAG repeat sizes, to the mediation effect. Furthermore, for each allele we divided the indirect effect by the total effect to obtain the index of mediation, i.e. the percentage of the effect of CAG repeat size variations on metabolites that is mediated by BMI.

7 FUNDING

This study was supported by a VENI-grant (#91615080) from the Netherlands Organization of Scientific Research. N.A.A. is partly supported by an Alzheimer's Association Research Grant (Award Number: AARG-19-616534) and a European Research Council Starting Grant (Number: 101041677). The NEO study is supported by the participating Departments, Division, and Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. DOM-K is supported by Dutch Science Organization (ZonMW-VENI Grant No. 916.14.023). T.O.F. was supported by the King Abdullah Scholarship Program and King Faisal Specialist Hospital & Research Center [No. 1012879283]. Funding for NESDA was obtained from the Netherlands Organization for Scientific Research (Geestkracht program grant 10-000-1002); the Center for Medical Systems Biology (CSMB, NWO Genomics), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL), VU University's Institutes for Health and Care Research (EMGO+) and Neuroscience Campus Amsterdam, University Medical Center Groningen, Leiden University Medical Center, National Institutes of Health (NIH, R01D0042157-01A, MH081802, Grand Opportunity grants 1RC2 MH089951 and 1RC2 MH089995). Part of the genotyping and analyses were funded by the Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health. Computing was supported by BiG Grid, the Dutch e-Science Grid, which is financially supported by NWO. The PROSPER study was supported by an investigator-initiated grant obtained from Bristol-Myers Squibb.

8 ACKNOWLEDGMENTS

The authors of the NEO study thank all participants, all participating general practitioners for inviting eligible participants, all research nurses for data collection, and the NEO study group: Pat van Beelen, Petra Noordijk, and Ingeborg de Jonge for coordination, laboratory, and data management. The authors are also thankful to Merel Boogaard for performing the genotyping assays.

8.1 Conflict of interest

R.L.-G. is a part-time clinical research consultant for Metabolon, Inc. All other co- authors have no conflicts of interest to declare.

9 AUTHOR CONTRIBUTIONS

T.O.F.- conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing-original draft. N.A.A.-, resources, funding acquisition, methodology, writing – review & editing. S.L.G.-, methodology, writing – review & editing. R.L.-G.- validation, writing – review & editing. R.d.M.- study design, conduct and data collection, resources, funding acquisition, writing – review & editing. Y.M.- project administration, resources, writing – review & editing. J.W.J.- resources, funding acquisition, writing – review & editing. F.R.R.- study design, funding acquisition, conceptualization. A.v.H.V. and K.W.v.D- conceptualization, supervision, writing – review & editing. D.O.M.-K.- conceptualization, supervision, funding acquisition, writing – review & editing.

10 REFERENCES

1 McColgan, P. and Tabrizi, S.J. (2018) Huntington's disease: a clinical review. *European Journal of Neurology*, **25**, 24-34.

2 Tabrizi, S.J., Flower, M.D., Ross, C.A. and Wild, E.J. (2020) Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities. *Nature Reviews Neurology*, **16**, 529-546.

3 Rawlins, M.D., Wexler, N.S., Wexler, A.R., Tabrizi, S.J., Douglas, I., Evans, S.J.W. and Smeeth, L. (2016) The Prevalence of Huntington's Disease. *Neuroepidemiology*, **46**, 144-153.

4 Caron, N.S., Wright, G.E.B. and Hayden, M.R., In *GeneReviews® [Internet]: Huntington Disease*. University of Washington, Seattle, WA, USA, in press.

5 Block, R.C., Dorsey, E.R., Beck, C.A., Brenna, J.T. and Shoulson, I. (2010) Altered cholesterol and fatty acid metabolism in Huntington disease. *J Clin Lipidol*, **4**, 17-23.

6 Aziz, N.A. and Roos, R.A. (2013) Characteristics, pathophysiology and clinical management of weight loss in Huntington's disease. **3**, 253-266.

7 Pringsheim, T., Wiltshire, K., Day, L., Dykeman, J., Steeves, T. and Jette, N. (2012) The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Movement Disorders*, **27**, 1083-1091.

8 Evans, S.J., Douglas, I., Rawlins, M.D., Wexler, N.S., Tabrizi, S.J. and Smeeth, L. (2013) Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *Journal of Neurology, Neurosurgery &amp; Psychiatry*, **84**, 1156-1160.

9 Gardiner, S.L., Boogaard, M.W., Trompet, S., de Mutsert, R., Rosendaal, F.R., Gussekloo, J., Jukema, J.W., Roos, R.A.C. and Aziz, N.A. (2019) Prevalence of Carriers of Intermediate and Pathological Polyglutamine Disease-Associated Alleles Among Large Population-Based Cohorts. *JAMA neurology*, **76**, 650-656.

10 Aziz, N.A., Van Der Burg, J.M.M., Landwehrmeyer, G.B., Brundin, P., Stijnen, T. and Roos, R.A.C. (2008) Weight loss in Huntington disease increases with higher CAG repeat number. *Neurology*, **71**, 1506-1513.

11 Leoni, V., Mariotti, C., Nanetti, L., Salvatore, E., Squitieri, F., Bentivoglio, A.R., Bandettini Del Poggio, M., Piacentini, S., Monza, D., Valenza, M. *et al.* (2011) Whole body cholesterol metabolism is impaired in Huntington's disease. *Neuroscience Letters*, **494**, 245-249.

12 Cheng, M.L., Chang, K.H., Wu, Y.R. and Chen, C.M. (2016) Metabolic disturbances in plasma as biomarkers for Huntington's disease. *J Nutr Biochem*, **31**, 38-44.

13 Mastrokolias, A., Pool, R., Mina, E., Hettne, K.M., van Duijn, E., van der Mast, R.C., van Ommen, G., t Hoen, P.A., Prehn, C., Adamski, J. *et al.* (2016) Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics*, **12**, 137.

14 Patassini, S., Begley, P., Reid, S.J., Xu, J., Church, S.J., Curtis, M., Dragunow, M., Waldvogel, H.J., Unwin, R.D., Snell, R.G. *et al.* (2015) Identification of elevated urea as a severe, ubiquitous metabolic defect in the brain of patients with Huntington's disease. *Biochem Biophys Res Commun*, **468**, 161-166.

15 Quintero Escobar, M., Pontes, J.G.D.M. and Tasic, L. (2021) Metabolomics in degenerative brain diseases. *Brain Research*, **1773**, 147704.

16 Stoy, N., Mackay, G.M., Forrest, C.M., Christofides, J., Egerton, M., Stone, T.W. and Darlington, L.G. (2005) Tryptophan metabolism and oxidative stress in patients with Huntington's disease. *Journal of Neurochemistry*, **93**, 611-623.

17 Gardiner, S.L., van Belzen, M.J., Boogaard, M.W., van Roon-Mom, W.M.C., Rozing, M.P., van Hemert, A.M., Smit, J.H., Beekman, A.T.F., van Grootheest, G., Schoevers, R.A. *et al.* (2017) Huntingtin gene repeat size variations affect risk of lifetime depression. *Translational psychiatry*, **7**, 1277.

18 Gardiner, S.L., Trompet, S., Sabayan, B., Boogaard, M.W., Jukema, J.W., Slagboom, P.E., Roos, R.A.C., van der Grond, J. and Aziz, N.A. (2019) Repeat variations in polyglutamine disease-associated genes and cognitive function in old age. *Neurobiology of aging*, **84**, 236.e217-236.e228.

19 de Mutsert, R., den Heijer, M., Rabelink, T.J., Smit, J.W., Romijn, J.A., Jukema, J.W., de Roos, A., Cobbaert, C.M., Kloppenburg, M., le Cessie, S. *et al.* (2013) The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. *Eur J Epidemiol*, **28**, 513-523.

20 Shepherd, J., Blauw, G.J., Murphy, M.B., Bollen, E.L., Buckley, B.M., Cobbe, S.M., Ford, I., Gaw, A., Hyland, M., Jukema, J.W. *et al.* (2002) Pravastatin in elderly individuals at risk of vascular disease (PROSPER): a randomised controlled trial. *Lancet*, **360**, 1623-1630.

21 Penninx, B.W., Beekman, A.T., Smit, J.H., Zitman, F.G., Nolen, W.A., Spinhoven, P., Cuijpers, P., De Jong, P.J., Van Marwijk, H.W., Assendelft, W.J. *et al.* (2008) The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int J Methods Psychiatr Res*, **17**, 121-140.

22 Albert, P.R. (2015) Why is depression more prevalent in women? *Journal of psychiatry & neuroscience : JPN*, **40**, 219-221.

23 Mäntyselkä, P., Kautiainen, H., Saltevo, J., Würtz, P., Soininen, P., Kangas, A.J., Ala-Korpela, M. and Vanhala, M. (2012) Weight change and lipoprotein particle concentration and particle size: A cohort study with 6.5-year follow-up. *Atherosclerosis*, **223**, 239-243.

24 Tikkanen, E., Jägerroos, V., Holmes, M.V., Sattar, N., Ala-Korpela, M., Jousilahti, P., Lundqvist, A., Perola, M., Salomaa, V. and Würtz, P. (2021) Metabolic Biomarker Discovery for Risk of Peripheral Artery Disease Compared With Coronary Artery Disease: Lipoprotein and Metabolite Profiling of 31 657 Individuals From 5 Prospective Cohorts. *Journal of the American Heart Association*, **10**, e021995.

25 Marston, N.A., Giugliano, R.P., Melloni, G.E.M., Park, J.G., Morrill, V., Blazing, M.A., Ference, B., Stein, E., Stroes, E.S., Braunwald, E. *et al.* (2022) Association of Apolipoprotein B-Containing Lipoproteins and Risk of Myocardial Infarction in Individuals With and Without Atherosclerosis: Distinguishing Between Particle Concentration, Type, and Content. *JAMA cardiology*, **7**, 250-256.

26 Guasch-Ferré, M., Hruby, A., Toledo, E., Clish, C.B., Martínez-González, M.A., Salas-Salvadó, J. and Hu, F.B. (2016) Metabolomics in Prediabetes and Diabetes: A Systematic Review and Meta-analysis. *Diabetes care*, **39**, 833-846.

27 Ahola-Olli, A.V., Mustelin, L., Kalimeri, M., Kettunen, J., Jokelainen, J., Auvinen, J., Puukka, K., Havulinna, A.S., Lehtimäki, T., Kähönen, M. *et al.* (2019) Circulating metabolites and the risk of type 2 diabetes: a prospective study of 11,896 young adults from four Finnish cohorts. *Diabetologia*, **62**, 2298-2309.

28 Cirulli, E.T., Guo, L., Leon Swisher, C., Shah, N., Huang, L., Napier, L.A., Kirkness, E.F., Spector, T.D., Caskey, C.T., Thorens, B. *et al.* (2019) Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metabolism*, **29**, 488-500.e482.

29 Kujala, U.M., Mäkinen, V.-P., Heinonen, I., Soininen, P., Kangas, A.J., Leskinen, T.H., Rahkila, P., Würtz, P., Kovanen, V., Cheng, S. *et al.* (2013) Long-term Leisure-time Physical Activity and Serum Metabolome. *Circulation*, **127**, 340-348.

30 Melkani, G.C. (2016) Huntington's Disease-Induced Cardiac Disorders Affect Multiple Cellular Pathways. *Reactive oxygen species (Apex, N.C.)*, **2**, 325-338.

31 Djoussé, L., Knowlton, B., Cupples, L.A., Marder, K., Shoulson, I. and Myers, R.H. (2002) Weight loss in early stage of Huntington's disease. *Neurology*, **59**, 1325-1330.

32 Gardiner, S.L., De Mutsert, R., Trompet, S., Boogaard, M.W., Van Dijk, K.W., Jukema, P.J.W., Slagboom, P.E., Roos, R.A.C., Pijl, H., Rosendaal, F.R. *et al.* (2019) Repeat length variations in polyglutamine disease-associated genes affect body mass index. *International Journal of Obesity*, **43**, 440-449.

33 Mastrokolias, A., Pool, R., Mina, E., Hettne, K.M., van Duijn, E., van der Mast, R.C., van Ommen, G., 't Hoen, P.A.C., Prehn, C., Adamski, J. *et al.* (2016) Integration of targeted metabolomics and transcriptomics identifies deregulation of phosphatidylcholine metabolism in Huntington's disease peripheral blood samples. *Metabolomics*, **12**, 137.

34 Delles, C., Rankin, N.J., Boachie, C., McConnachie, A., Ford, I., Kangas, A., Soininen, P., Trompet, S., Mooijaart, S.P., Jukema, J.W. *et al.* (2018) Nuclear magnetic resonance-based metabolomics identifies phenylalanine as a novel predictor of incident heart failure hospitalisation: results from PROSPER and FINRISK 1997. *Eur J Heart Fail*, **20**, 663-673.

35 de Kluiver, H., Jansen, R., Milaneschi, Y., Bot, M., Giltay, E.J., Schoevers, R. and Penninx, B.W.J.H. (2021) Metabolomic profiles discriminating anxiety from depression. *Acta Psychiatr Scand*, **144**, 178-193.

36 Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*, **13**, 36-46.

37 Soininen, P., Kangas, A.J., Würtz, P., Suna, T. and Ala-Korpela, M. (2015) Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. **8**, 192-206.

38 Joshi, R., Wannamethee, G., Engmann, J., Gaunt, T., Lawlor, D.A., Price, J., Papacosta, O., Shah, T., Tillin, T., Whincup, P. *et al.* (2021) Establishing reference intervals for triglyceride-containing lipoprotein subfraction metabolites measured using nuclear magnetic resonance spectroscopy in a UK population. *Annals of clinical biochemistry*, **58**, 47-53.

39 Soininen, P., Kangas, A.J., Wurtz, P., Suna, T. and Ala-Korpela, M. (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet*, **8**, 192-206.

40 Faquih, T. (2020). Zenodo, Zenodo, Vol. 2020.

41 Faquih, T., van Smeden, M., Luo, J., le Cessie, S., Kastenmüller, G., Krumsiek, J., Noordam, R., van Heemst, D., Rosendaal, F.R., van Hylckama Vlieg, A. *et al.* (2020) A Workflow for Missing Values Imputation of Untargeted Metabolomics Data. *Metabolites*, **10**.

42 Aziz, N.A., Jurgens, C.K., Landwehrmeyer, G.B., van Roon-Mom, W.M., van Ommen, G.J., Stijnen, T. and Roos, R.A. (2009) Normal and mutant HTT interact to affect clinical severity and progression in Huntington disease. *Neurology*, **73**, 1280-1285.

43 Gardiner, S.L., de Mutsert, R., Trompet, S., Boogaard, M.W., van Dijk, K.W., Jukema, P.J.W., Slagboom, P.E., Roos, R.A.C., Pijl, H., Rosendaal, F.R. *et al.* (2019) Repeat length variations in polyglutamine disease-associated genes affect body mass index. *International journal of obesity (2005)*, **43**, 440-449.

44 R Core Team. (2019). R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. in press.

45 Lee M. A, M.O., Hughes D, Wade K. H, Corbin L. J, McGuinness L. J, Timpson N. J. (2020), in press.

46 Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, **32**, 2847-2849.

47 Gu, Z., Gu, L., Eils, R., Schlesner, M. and Brors, B. (2014) circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England)*, **30**, 2811-2812.

48 StataCorp. (2019). Stata Press, in press., pp. 475-563.

49 Rücker, G. and Schwarzer, G. (2014) Presenting simulation results in a nested loop plot. *BMC Medical Research Methodology*, **14**, 129.

50 Kammer, M. (2022), in press.

51 Baron, R.M. and Kenny, D.A. (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**, 1173-1182.

52 Aroian, L.A. (1947) The Probability Function of the Product of Two Normally Distributed Variables. *The Annals of Mathematical Statistics*, **18**, 265-271, 267.

53 MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G. and Sheets, V. (2002) A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods*, **7**, 83-104.

11 FIGURE LEGENDS

Figure 1: Flow chart of the exclusion criteria in the NEO, NESDA, and PROSPER studies before and after pooling.

Figure 2: Mean centered CAG repeat size distribution in *HTT* alleles in the pooled and individual datasets.

Figure 3: Circular plots for the 145 metabolites concentrations associated with the larger CAG repeat size in the long *HTT* allele. Each dot represents the effect estimates for the log-transformed metabolite levels. The lines crossing the circles represent the 95% confidence intervals of the estimates. Filled circles denote statistically significant estimates after adjustment for multiple testing (i.e., p<0.00145). The outer numbered rings represent the different metabolite groups.

Figure 4: Estimation of metabolite levels related to VLDL (A), IDL (B), LDL (C), HDL (D), and other metabolites (E) at 15, 20 and 35 CAG repeat sizes.

12 TABLES

Table 1. Baseline characteristics of the three included studies.

	NEO*	**PROSPER**	**NESDA**	**Overall**
N	4,510	4,035	1,712	10,257
Age in years (SD)	55.93 (5.9)	75.79 (3.4)	42.44 (12.9)	61.50 (14.2)
Sex = Female (%)	2,378 (52.7)	2,079 (51.5)	1,129 (65.9)	5,586 (54.5)
Country (%)				
Scotland	0	1,808 (44.8)	0	1,808 (17.6)
Ireland	0	1,448 (35.9)	0	1,448 (14.1)
Netherlands	4,510 (100.0)	779 (19.3)	1,712 (100.0)	7,001 (68.3)
BMI (SD)	26.3 (3.50)	26.8 (4.1)	25.5 (5.0)	28.0 (4.9)
CAG repeats size (median [range])				
HTT Short allele	17 [9, 26]	17 [9, 26]	17 [9, 26]	17 [9, 26]
HTT Long allele	19 [15, 35]	19 [15, 35]	19 [15, 35]	19 [15, 35]

* Means and percentages were weighted to the BMI distribution of the Dutch general.

Table 2: Largest mediation estimates by BMI in the associations between the *HTT* CAG repeats and metabolite levels.

Metabolite	Mediation Estimate	Direct Effect Estimate	Total Effect Estimate	Sobel's Test P-value
Phospholipids in very large HDL	0.0013	-0.0031	-0.0017	0.0060
Free cholesterol in very large HDL	0.0011	-0.0032	-0.0020	0.0063
Total lipids in very large HDL	0.0010	-0.0027	-0.0017	0.0064
Concentration of very large HDL particles	0.0010	-0.0027	-0.0016	0.0063
Total cholesterol in very large HDL	0.0009	-0.0025	-0.0016	0.0069
Cholesterol esters in very large HDL	0.0009	-0.0024	-0.0015	0.0072
Phospholipids in XL VLDL	-0.0021	0.0052	0.0031	0.0060
Cholesterol esters in XL VLDL	-0.0018	0.0060	0.0042	0.0071

13 ABBREVIATIONS

HD: Huntington disease; *HTT*: huntingtin gene; BMI: body mass index; CAG: cytosine-adenine-guanine; CAD: coronary artery disease; PAD: peripheral artery disease; XL-VLDL: extremely large very low density lipoprotein; VLDL: very low density lipoprotein; IDL: intermediate density lipoprotein; LDL: low density lipoprotein; HDL: high density lipoprotein; NEO: Netherlands Epidemiology of Obesity; PROSPER: Prospective Study of Pravastatin in the Elderly at Risk; NESDA: Netherlands Study of Depression and Anxiety; apoB: apolipoprotein B; HDL3: high-density lipoprotein 3 cholesterol; PCR: multiplex polymerase chain reaction.