Genome-wide association study of habitual physical activity in over 277,000 UK Biobank participants identifies novel variants and genetic correlations with chronotype and obesity-related traits.

Short Title: Genome-wide association study of habitual physical activity

Yann C. Klimentidis[1*], David A. Raichlen[2], Jennifer Bea[3,4], David O. Garcia[5], Lawrence J. Mandarino[6], Gene E. Alexander[7], Zhao Chen[1], Scott B. Going[4]

[1] Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, USA

[2] School of Anthropology, University of Arizona, Tucson, Arizona, USA

[3] Department of Medicine, University of Arizona, Tucson, Arizona, USA

[4] Department of Nutritional Sciences, University of Arizona, Tucson, Arizona, USA

[5] Department of Health Promotion Sciences, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, USA

[6] Center for Disparities in Diabetes, Obesity and Metabolism, Division of Endocrinology, Diabetes and Metabolism, Department of Medicine, University of Arizona, Tucson, Arizona, USA

[7] Departments of Psychology and Psychiatry, Neuroscience and Physiological Sciences Interdisciplinary Programs, BIO5 Institute, and Evelyn F. McKnight Brain Institute, University of Arizona, Tucson, Arizona, USA; Arizona Alzheimer's Consortium, Phoenix, Arizona, USA

*** Corresponding Author:** Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, Arizona, 85724, 520-621-0147, yann@email.arizona.edu

Abstract

Physical activity (PA) protects against a wide range of diseases. Engagement in habitual PA has been shown to be heritable, motivating the search for specific genetic variants that explain variation in habitual PA and may ultimately improve efforts to promote PA and target the best type of PA for each individual. We used data from the UK Biobank to perform the largest genome-wide association study of PA, using four measures based on self-report (n=277,656) and accelerometry (n=67,808). Replication was then sought in the Atherosclerosis Risk in Communities (ARIC) study (n=8,556). In the UK Biobank, we identified 17 genome-wide loci across the four PA measures. Interestingly, rs429358 of the *APOE* gene was the most strongly associated variant with any single PA measure and was at least nominally associated with three of the four PA measures examined. We also identified three loci (*DNAJC1*, *DCAF5*, and *PML*) consistently associated with PA across all four measures. Tissue enrichment analyses implicate the brain and pituitary gland as locations where PA-associated loci may exert their actions. Genetic correlation analyses suggest a positive genetic correlation of PA with early-morning chronotype and psychiatric traits, and a negative genetic correlation of PA with obesity-related traits. Using data from the GIANT consortium, we identify several loci that are associated with both increased waist circumference and decreased PA. Although very small effect sizes precluded replication of individual loci in ARIC, we found consistent overall genetic correlations of PA with other traits. These results provide new insight into the genetic basis of habitual PA, and the genetic links connecting PA and obesity.

Introduction

A physically active lifestyle has been shown to protect against a wide range of diseases, including cardiovascular disease, cancer, type-2 diabetes, and osteoporosis [1–4]. Levels of engagement in physical activity (PA) vary across individuals, and many people do not meet recommended levels to achieve health benefits. Although cultural, economic, and other environmental factors influence PA engagement [5,6], genetic factors also likely play a role. Understanding the genetic factors underlying inter-individual variation will better inform efforts to promote PA and potentially allow targeting the best type of PA for each person, what might be called "Precision Exercise Prescription".

Evidence of genetic factors underlying the propensity to exercise in humans has been demonstrated in a number of studies [7–14]. Several studies have utilized a candidate gene approach to identify specific genetic variants associated with a proclivity towards PA [8,15–19]. These studies generally focused on genes related to the serotonin and dopamine systems, energy metabolism, and neurotrophic factors. However, to our knowledge there have been only two previous reports of genome-wide association studies (GWAS) of PA [20,21], neither of which identified a locus at genome-wide significance, likely due to relatively small sample sizes. Thus, while previous work strongly suggests a genetic basis for engagement in PA, the genes that contribute strongly to this healthy lifestyle remain unknown.

In this study, we conduct the largest GWAS of PA to date, aiming to identify genetic variants associated with self-reported and accelerometry-based levels of habitual, leisure-time PA. We sought to identify variants in the UK Biobank, a large cohort study of 500,000 adults measured across a wide range of characteristics including genome-wide markers. We then examined the genetic correlation of PA with other traits, examined putative tissues where PA genes may exert their effects, and attempted to replicate our identified loci using self-reported PA in the Atherosclerosis Risk in Communities (ARIC) study.

Methods

Studies

Data from the UK Biobank study were used for discovery of variants. Briefly, the UK Biobank is a large prospective cohort study of approximately a half-million adults (ages 40-69) living in the United Kingdom (UK), recruited from 22 centers across the UK [22]. All participants provided written informed consent. For replication we used data from the ARIC study, which is a prospective cohort study of over 15,000 adults aged 45-64 years that took place in four United States communities. Details of the ARIC study can be found elsewhere [23]. To reduce the potential for confounding by population stratification, we included only individuals of white race/ethnicity in both studies.

Physical activity

In the UK Biobank, self-reported levels of physical activity during work and leisure time were measured via a touchscreen questionnaire, in a fashion similar to the International Physical Activity Questionnaire [24]. For moderate PA (MPA), participants were asked: "In a typical WEEK, on how many days did you do 10 minutes or more of moderate physical activities like carrying light loads, cycling at normal pace? (Do not include walking)". For vigorous PA (VPA), participants were asked: "In a typical WEEK, how many days did you do 10 minutes or more of vigorous physical activity? (These are activities that make you sweat or breathe hard such as fast cycling, aerobics, heavy lifting)". For each of these questions, those who indicated 1 or more such days were then asked "How many minutes did you usually spend doing moderate/vigorous activities on a typical DAY". Participants were asked to include activities performed for work, leisure, travel and around the house. We excluded individuals who selected "prefer not to answer" or "do not know" on the above questions, those reporting not being able to walk, and individuals reporting more than 16 hours of either MPA or VPA per day. Those reporting >3hr/day of VPA or MPA were recoded to 3 hours, as recommended [25]. Moderate-to-vigorous PA (MVPA) was

88 calculated by taking the sum of total minutes/week of MPA multiplied by four and the total number of
89 VPA minutes/week multiplied by eight, corresponding to their metabolic equivalents, as previously
90 described [24,26].

91 Since heritability has previously been shown to be higher for intense/vigorous physical activity
92 [13], we also considered VPA on its own. Because the distribution of minutes/week of VPA was highly
93 skewed and zero-inflated, we chose to dichotomize minutes/week of VPA into those who reported 0 days
94 of VPA, and those reporting 3 or more days of VPA and also reporting a typical duration of VPA that is
95 25 minutes or greater.

96 Also, in the UK Biobank, approximately 100,000 participants wore an Axivity AX3 wrist-worn
97 accelerometer, as previously described [27]. We examined two measures derived from up to seven days
98 of accelerometer wear: overall acceleration average, and fraction of accelerations > 425 milli-gravities
99 (mg) [27]. Since the variable that is available in the UK Biobank is the fraction < 425 mg, we subtracted 1
100 from this variable. The 425 mg cutoff was chosen because this corresponds to an equivalent of vigorous
101 physical activity, as previously reported [28]. For both accelerometry variables, we excluded individuals
102 with less than three days (72 hours) of data, those not having data in each one-hour period of the 24-hour
103 cycle, and outliers with values more than 4 standard deviations above the mean.

104 For the replication study (ARIC), self-reported PA was assessed for sports/exercise, within the
105 previous year, based on a modification of the Baecke questionnaire [29,30]. The sport/exercise index is
106 based on responses to 4 items: frequency of participation in sports/exercise; frequency of sweating during
107 sports/exercise; a subjective rating of the frequency of participation in sports/exercise compared to others
108 in the same age group; the sum of frequency, duration, and intensity of up to 4 sports/exercises.

109

110 *Genotypes*

111 The majority of UK Biobank participants were genotyped with the Affymetrix UK Biobank
112 Axiom Array (Santa Clara, CA, USA), while 10% of participants were genotyped with the Affymetrix
113 UK BiLEVE Axiom Array. Detailed quality control and imputation procedures are described elsewhere
114 [31]. Imputation was performed using a combined panel of the Haplotype Reference Consortium [32] and
115 the UK10K haplotype resource [33]. Since corrections for potential problems with the position
116 assignment of the SNPs from the UK10K haplotype resource were not available at the time of analysis,
117 we only included SNPs imputed from the Haplotype Reference Consortium. To minimize the possibility
118 of confounding due to population stratification, only participants who self-identified as White/British
119 (European-descent), and who were not population outliers based on principal components analysis (PCA)
120 were included. Individuals were excluded based on unusually high heterozygosity or >5% missing rate, a
121 mismatch between self-reported and genetically-inferred sex, and on relatedness (i.e. samples not used in
122 primary PCA). SNP exclusions were made based on Hardy-Weinberg equilibrium ($p<1x10^{-6}$), high
123 missingness (>1.5%), low minor allele frequency (<0.1%), and low imputation quality (info<0.4).

124 In ARIC, participants were genotyped with the Affymetrix Genome-Wide Human SNP Array 6.0
125 (Affymetrix, Santa Clara, CA, USA). Standard quality control procedures were implemented prior to
126 imputation with IMPUTE2 [34], using all individuals in the 1,000 Genomes phase 1 integrated v3 release.
127 Quality-control procedures consisted of excluding SNPs with minor allele frequency < 1%, with
128 missingness >10%, and SNPs out of Hardy-Weinberg equilibrium ($p<1 \times 10^{-6}$), and excluding individuals
129 with SNP missingness > 10%. We used principal components for the European-ancestry group as
130 provided by ARIC in dbGaP. Briefly, LD pruning resulted in 71,702 SNPs that were used to derive
131 principal components. A total sample size of 8,556 participants was used in the analysis.

132

133 *Statistical analyses*

For the continuous variables in the UK Biobank (MVPA and accelerometry variables) we created an adjusted phenotype corresponding to the residual of the regression of the following independent variables on the respective dependent PA variable: age, sex, genotyping chip, first ten genomic principal components, center, season (month) at center visit or wearing accelerometer (coded 0 for Winter, 1 for Fall or Spring, and 2 for Summer). In sensitivity analyses, we considered the inclusion of the additional following covariates: levels of physical activity at work (coded as 0 by default, 1 for 'sometimes', 2 for 'usually', and 3 for 'always'), extent of walking or standing at work (coded similarly as previous variable), and the Townsend Deprivation Index (TDI; a composite measure of deprivation as previously described [35,36]). We also considered a third model in which body mass index (BMI) was included as an additional covariate. Since the MVPA and fraction of accelerations > 425 mg variables exhibited skewed distributions, we inverse-normalized these variables prior to inclusion in the models. Model residuals conformed to the assumptions of normality and homoscedasticity. For the categorical variables, we used logistic regression, including the same set of covariates listed above. We conducted two additional analyses. First, we sought to identify variants consistently associated with PA across all four measures. We thus searched for variants associated in the same direction, with $p<0.001$ with each of the four PA phenotypes. The prior probability under the null hypothesis for a SNP association at this level with all four phenotypes and with the same direction of association is $0.001 \times 0.0005 \times 0.0005 \times 0.0005 = 1.25 \times 10^{-13}$. However, it is important to note that these four phenotypes are correlated with each other. Second, given the observed strong and consistent negative genetic correlation of all PA measures with waist circumference we used summary statistics from the GIANT consortium [37] to identify loci that are consistently associated with all four measures of PA in the same direction ($p<0.01$ in each of four PA measures) and with waist circumference in the opposite direction ($p<0.01$).

Given the association that we identified with the rs428358 variant in *APOE*, we examined the association of the *APOE* ε4 genotype, using this SNP along with the rs7412 SNP. Individuals with homozygous CC genotypes at both of these SNPs were classified as homozygous for the APOE ε4 allele.

159 Individuals with homozygous CC genotypes at either SNP and heterozygous at the other SNP were
160 classified as being heterozygous for the ε4 allele. We excluded individuals heterozygous at both SNPs.
161 We assumed an additive model in association testing.

162 All genome-wide significant loci were examined in ARIC, where we modeled PA as a continuous
163 variable (as described above). We used multiple linear regression to model PA as a function of: age, sex,
164 first ten genomic principal components, BMI, center, season (coded in the same way as described above).
165 Residuals from this model conformed to the assumptions of normality and homoscedasticity, and were
166 used as the outcome in the genome-wide SNP association analysis. A replication was considered
167 successful if the p-value in ARIC was <0.0025, corresponding to a Bonferroni correction for 20 tests
168 performed. All GWAS analyses were performed with PLINK 2.0, assuming an additive genetic model
169 [38]. Additional analyses were performed with R statistical software[39]. To examine association of
170 genetic variants with gene expression patterns in different tissues, we used the web-based platform,
171 Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS) [40], which
172 uses data from GTEx [41].

173 We used the summary statistics from our GWAS to examine genetic correlation of PA with over
174 200 traits and diseases using LD score regression [42,43], implemented in an online interface
175 (http://ldsc.broadinstitute.org/). The genetics of these other traits and diseases are inferred from
176 previously published GWAS. A significant genetic correlation was considered to have a $p < 2.5 \times 10^{-4}$,
177 assuming a correction for 200 different tests, which is conservative given that many of the traits/diseases
178 tested are correlated with each other. LD score regression intercepts and chip heritability estimates were
179 also recorded.

180

181 **Results**

182 *Self-reported PA in UK Biobank*

183 A summary of self-report PA variables can be found in Table 1. BMI and TDI were consistently
184 negatively associated with these variables, whereas season and physical activity at work were consistently
185 positively associated with them (see Supplementary Table 1). Self-report PA measures were weakly
186 correlated with accelerometry-based measures (see Supplementary Table 2). 'Chip heritability' estimates
187 for both self-report PA measures were approximately 5% (Supplementary Table 3). LD score regression
188 intercepts (<1.02) suggest no significant systematic inflation of test statistics. In over 277,000 individuals,
189 we find ten loci significantly associated ($p < 5 \times 10^{-8}$) with MVPA and three loci with VPA (see Figure 1
190 and Table 2). Most notably, the C allele at SNP rs429358 in *APOE* is associated with higher self-reported
191 MVPA. This allele is also nominally associated with higher levels of other measures of PA (VPA: $p=8.4$
192 $\times 10^{-6}$; average acceleration: $p=3.7 \times 10^{-3}$). Testing the association of the Alzheimer's disease-associated
193 *APOE* ε4 allele with MVPA resulted in nearly identical findings. In models adjusted for other covariates,
194 this *APOE* variant remained genome-wide significant (see Supplementary Tables 4 and 5). We did not
195 observe any loci significantly associated with both MVPA and VPA. In models adjusting for additional
196 covariates, the results were very similar with slightly fewer loci reaching genome-wide significance for
197 MVPA, and slightly more for VPA (see Supplementary Tables 4 and 5 and Supplementary Figures 1 and
198 2). In ARIC, we observed no significant replication of any variant. However, we note similar direction of
199 association for 8 of the 13 variants, including the *APOE* SNP (see Table 2).

200

201 *Accelerometer-based PA in UK Biobank*

202 'Chip heritability' estimates for the accelerometry-based measures were higher (15% for average
203 acceleration, and 11% for fraction of accelerations >425 mg) than for self-report PA measures
204 (Supplementary Table 3). LD score regression intercepts (<1.009) suggest no significant systematic
205 inflation of test statistics. Three loci were found to be significantly associated with average acceleration
206 and one locus with fraction of accelerations >425 mg (see Table 2 and Figure 1).

207 We found three loci associated with each of the four PA measures at $p<0.001$, and with the same
208 direction of association (Table 3). These include an intronic variant in *DNAJC1*, a variant just upstream of
209 *DCAF5*, and a missense variant in *PML*. No SNP was successfully replicated in ARIC.

210

211 *Genetic correlation and tissue enrichment analysis*

212 We examined the genetic correlation between habitual PA and over 200 hundred other
213 traits/diseases. Most notably, we find highly significant negative genetic correlations of PA with waist
214 and hip circumference, body fat, obesity, BMI, and triglycerides (see Figure 2). This negative genetic
215 correlation with obesity and related traits is particularly accentuated for the accelerometry-derived PA
216 variables (see Figure 2). We found three potentially pleiotropic loci associated with both lower physical
217 activity and greater waist circumference (see Table 4), including the *DNAJC1* variant previously
218 identified as being consistently associated with PA. We find a positive genetic correlation with
219 chronotype (morning-type) and psychiatric traits, and a negative genetic correlation of self-reported PA
220 with educational attainment traits. Results remained very similar in GWAS models adjusted for activity at
221 work and TDI variables (see Supplementary Figure 3), except that the negative correlation with
222 educational attainment traits exhibited a reversed direction of association. Similarly, upon addition of
223 BMI as a covariate, the direction of genetic correlation between PA and obesity traits was reversed (see
224 Supplementary Figure 4). As we note below, caution is warranted in interpreting results from these
225 adjusted models. In ARIC, we found similar genetic correlations of PA with waist circumference (r_g=-
226 0.13), chronotype (r_g=0.04), and schizophrenia (r_g=0.32), although these do not reach statistical
227 significance.

228 Tissue enrichment analysis using eQTL data from GTEx implicate the brain and pituitary as
229 primary tissues through which the PA-associated loci may exert their effects (see Figure 3 and
230 Supplementary Figure 5).

231

232 **Discussion**

233 Given the importance of physical activity for multiple dimensions of health, and its' reported
234 heritability, we sought to identify genetic variants that are associated with engagement in habitual
235 physical activity, while considering important covariates such as season, physical activity at work, socio-
236 economic status, and BMI.

237 In the UK Biobank, with a very large sample size and multiple measures of PA, we identified
238 several genome-wide significant SNPs. Interestingly, a well-established variant in *APOE* (part of *APOE*
239 ε4 allele), strongly implicated in Alzheimer's disease [44,45], exhibited one of the strongest associations
240 with PA. How the *APOE* risk allele is associated with greater physical activity is not clear. An exercise
241 training study found that *APOE* ε4 carriers had a greater increase in aerobic capacity [46]. This increased
242 capacity could reinforce engagement in PA. Another potential explanation for our finding is that
243 individuals with a known family history of dementia or Alzheimer's disease purposefully increase their
244 levels of PA in order to avoid succumbing to these diseases. Nevertheless, it is important to note that this
245 association may lead to spurious gene-environment interactions [47]. This is particularly pertinent in the
246 case of the reported interactions of *APOE* with physical activity that suggest PA can have a beneficial
247 effect on Alzheimer's disease risk [e.g. 44–46]. Although this association remained after adjustment for
248 BMI and other covariates, it remains to be determined whether some residual confounding exists. None of
249 the identified loci were successfully replicated in the ARIC data, likely due to the very small effect sizes
250 observed in the UK Biobank, the smaller sample size in ARIC, and the different methods of measuring
251 physical activity in these studies. Additionally, it is important to note that none of the identified loci have
252 previously been found to be associated with PA or related traits [e.g. 16,17].

253 Previous studies have shown that BMI-associated genetic variants are also associated with PA
254 [51,52]. Similarly, we find an overall shared genetic basis for PA with several obesity-related traits,

255 suggesting that genetic risk for obesity coincides with genetic propensity for lower PA. There is likely a
256 complex set of genetic, environmental, and phenotypic factors that connect PA and obesity across the
257 lifespan, that involve many pleiotropic genetic factors. Although the direction of correlation is reversed
258 when BMI is included as a covariate, caution is warranted in interpreting results of genetic associations in
259 which heritable covariates are included in the association model [53]. Another major potential source of
260 bias is collider bias, which occurs when one controls for a variable (i.e. BMI) that is caused by both
261 another covariate (i.e. gene) and the outcome variable in the model (i.e. PA) [54,55].

262 Our study is strengthened by the large sample size, the availability of both self-reported and
263 accelerometer-based measures of PA, and the availability of a replication cohort from a different country.
264 However, we note several limitations. Given the relatively small genetic effect sizes observed for these
265 PA phenotypes, we were insufficiently powered to replicate associations in the much smaller sample size
266 in ARIC. Additional replication studies are thus needed to more robustly identify PA-associated loci.
267 Furthermore, the self-report measure of PA used in ARIC differed from the one used in the UK Biobank.
268 Both self-reported and accelerometer-based measures of PA are subject to various biases. Since both the
269 UK Biobank and ARIC cohorts are comprised of middle- to late-middle-aged adults, the extent to which
270 these results generalize to other age groups is not known. Furthermore, our results may not generalize to
271 other ethnic/racial groups.

272 In conclusion, our study revealed several important new findings. Effect sizes were generally
273 very small, given the very large sample size, variant allele frequencies, and modest p-values. *APOE*, a
274 major risk factor for Alzheimer's disease, is associated with higher levels of PA, which could importantly
275 lead to spurious *APOE*-by-PA interactions in epidemiological studies. We found a negative genetic
276 correlation of PA with obesity and related traits, suggesting extensive pleiotropy and complex
277 connections among PA and obesity. The identification of genetic factors that predispose to high or low
278 levels of physical activity will lead to a better understanding of the biological mechanisms underlying
279 these proclivities. It may also lead to the identification of individuals less likely to engage in and/or

adhere to physical activity, and consequently to the development of tailored behavioral strategies. Finally, the integration of genetic information with lifestyle and environmental information may point to how lifestyle/environmental factors interact with genetic factors to influence levels of PA.

Acknowledgments

This research was conducted using the UK Biobank Resource under Application Number 15678. We thank the participants and organizers of the UK Biobank. We also thank the participants and organizers of the ARIC study. Data from ARIC was obtained from dbGaP through accession number phs000280.v2.p1. The authors would like to acknowledge support from the National Institute of Diabetes and Digestive and Kidney Diseases grant (K01DK095032), the National Institute on Aging (AG019610), the State of Arizona and Arizona Department of Health Services (ADHS), and the McKnight Brain Research Foundation.

Atherosclerosis Risk in Communities

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Funding for GENEVA was provided by National Human Genome Research Institute grant U01HG004402 (E. Boerwinkle). The authors thank the staff and participants of the ARIC study for their important contributions.

LDHUB Acknowledgements

301 We gratefully acknowledge all the studies and databases that made GWAS summary data available:
302 ADIPOGen (Adiponectin genetics consortium), C4D (Coronary Artery Disease Genetics Consortium),
303 CARDIoGRAM (Coronary ARtery DIsease Genome wide Replication and Meta-analysis), CKDGen
304 (Chronic Kidney Disease Genetics consortium), dbGAP (database of Genotypes and Phenotypes),
305 DIAGRAM (DIAbetes Genetics Replication And Meta-analysis), ENIGMA (Enhancing Neuro Imaging
306 Genetics through Meta Analysis), EAGLE (EArly Genetics & Lifecourse Epidemiology Eczema
307 Consortium, excluding 23andMe), EGG (Early Growth Genetics Consortium), GABRIEL (A
308 Multidisciplinary Study to Identify the Genetic and Environmental Causes of Asthma in the European
309 Community), GCAN (Genetic Consortium for Anorexia Nervosa), GEFOS (GEnetic Factors for
310 OSteoporosis Consortium), GIANT (Genetic Investigation of ANthropometric Traits), GIS (Genetics of
311 Iron Status consortium), GLGC (Global Lipids Genetics Consortium), GPC (Genetics of Personality
312 Consortium), GUGC (Global Urate and Gout consortium), HaemGen (haemotological and platelet traits
313 genetics consortium), HRgene (Heart Rate consortium), IIBDGC (International Inflammatory Bowel
314 Disease Genetics Consortium), ILCCO (International Lung Cancer Consortium), IMSGC (International
315 Multiple Sclerosis Genetic Consortium), MAGIC (Meta-Analyses of Glucose and Insulin-related traits
316 Consortium), MESA (Multi-Ethnic Study of Atherosclerosis), PGC (Psychiatric Genomics Consortium),
317 Project MinE consortium, ReproGen (Reproductive Genetics Consortium), SSGAC (Social Science
318 Genetics Association Consortium) and TAG (Tobacco and Genetics Consortium), TRICL
319 (Transdisciplinary Research in Cancer of the Lung consortium), UK Biobank. We gratefully acknowledge
320 the contributions of Alkes Price (the systemic lupus erythematosus GWAS and primary biliary cirrhosis
321 GWAS) and Johannes Kettunen (lipids metabolites GWAS).

322

323 **References**

324 1. Fiuza-Luces C, Garatachea N, Berger NA, Lucia A (2013) Exercise is the real polypill. Physiol.

325 2. US Surgeon General (1996) Physical Activity and Health: A Report of the Surgeon General. S/N
326 017-023-00196-5.

327 3. Blair SN (2009) Physical inactivity: the biggest public health problem of the 21st century. Br J
328 Sports Med 43: 1 LP-2.

329 4. Kaplan GA, Strawbridge WJ, Cohen RD, Hungerford LR (1996) Natural history of leisure-time
330 physical activity and its correlates: associations with mortality from all causes and cardiovascular
331 disease over 28 years. Am J Epidemiol.

332 5. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJF, et al. (2012) Correlates of physical activity:
333 Why are some people physically active and others not? Lancet 380: 258–271.

334 6. Trost SG, Owen N, Bauman AE, Sallis JF, Brown W (2002) Correlates of adults' participation in
335 physical activity: review and update. Med Sci Sport Exerc.

336 7. den Hoed M, Brage S, Zhao JH, Westgate K, Nessa A, et al. (2013) Heritability of objectively
337 assessed daily physical activity and sedentary behavior. Am J Clin Nutr 98: 1317–1325.

338 8. Gielen M, Westerterp-Plantenga MS, Bouwman FG, Joosen AMCP, Vlietinck R, et al. (2014)
339 Heritability and genetic etiology of habitual physical activity: a twin study with objective
340 measures. Genes Nutr 9: 415.

341 9. Stubbe JH, Boomsma DI, Vink JM, Cornes BK, Martin NG, et al. (2006) Genetic influences on
342 exercise participation in 37.051 twin pairs from seven countries. PLoS One 1.

343 10. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, et al. (2003) Heritability of
344 adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6: 399–408.

345 11. Joosen AMCP, Gielen M, Vlietinck R, Westerterp KR (2005) Genetic analysis of physical activity
346 in twins. Am J Clin Nutr 82: 1253–1259.

347 12. Pérusse L, Tremblay A, Leblanc C, Bouchard C (1989) Genetic and environmental influences on
348 level of habitual physical activity and exercise participation. Am J Epidemiol.

349 13. Lauderdale DS, Fabsitz R, Meyer JM, Sholinsky P, Ramakrishnan V, et al. (1997) Familial
350 determinants of moderate and intense physical activity: a twin study. Med Sci Sports Exerc 29:
351 1062–1068.

352 14. Kaprio J, Koskenvuo M, Sarna S (1981) Cigarette smoking, use of alcohol, and leisure-time
353 physical activity among same-sexed adult male twins. Prog Clin Biol Res 69 Pt C: 37–46.

354 15. Thompson PD, Tsongalis GJ, Ordovas JM, Seip RL, Bilbie C, et al. (2006) Angiotensin-
355 converting enzyme genotype and adherence to aerobic exercise training. Prev Cardiol.

356 16. Herring MP, Sailors MH, Bray MS (2014) Genetic factors in exercise adoption, adherence and
357 obesity. Obes Rev 15: 29–39.

358 17. Wilkinson A V., Gabriel KP, Wang J, Bondy ML, Dong Q, et al. (2013) Sensation-seeking genes
359 and physical activity in youth. Genes, Brain Behav.

360 18. Caldwell Hooper AE, Bryan AD, Hagger MS (2014) What keeps a body moving? The brain-
361 derived neurotrophic factor val66met polymorphism and intrinsic motivation to exercise in
362 humans. J Behav Med 37: 1180–1192.

363 19. Lightfoot JT (2011) Current understanding of the genetic basis for physical activity. J Nutr 141:
364 526–530.

365 20. De Moor MHM, Liu Y-J, Boomsma DI, Li J, Hamilton JJ, et al. (2009) Genome-wide association
366 study of exercise behavior in Dutch and American adults. Med Sci Sports Exerc 41: 1887–1895.

367 21. Kim J, Oh S, Min H, Kim Y, Park T (2011) Practical issues in genome-wide association studies
368 for physical activity. Ann N Y Acad Sci 1229: 38–44.

369 22. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. (2015) UK Biobank: An Open Access
370 Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old
371 Age. PLoS Med 12.

372 23. The Aric Investigators (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design
373 and objectives. The ARIC investigators. Am J Epidemiol 129: 687–702.

374 24. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, et al. (2003) International physical
375 activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 35: 1381–1395.

376 25. Ipaq, Sci M, Med S, Hardcastle S, Blake N, et al. (2002) Guidelines for Data Processing and
377 Analysis of the International Physical Activity Questionnaire (IPAQ) - Short Form. J Am Diet
378 Assoc.

379 26. Ekelund U, Sepp H, Brage S, Becker W, Jakes R, et al. (2006) Criterion-related validity of the last
380 7-day, short form of the International Physical Activity Questionnaire in Swedish adults. Public
381 Health Nutr.

382 27. Doherty A, Jackson D, Hammerla N, Plötz T, Olivier P, et al. (2017) Large Scale Population
383 Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study.
384 PLoS One 12: e0169649.

385 28. Hildebrand M, Van Hees VT, Hansen BH, Ekelund U (2014) Age group comparability of raw
386 accelerometer output from wrist-and hip-worn monitors. Med Sci Sports Exerc.

387 29. Baecke JA, Burema J, Frijters JE (1982) A short questionnaire for the measurement of habitual
388 physical activity in epidemiological studies. Am J Clin Nutr 36: 936–942.

389 30. Folsom AR, Arnett DK, Hutchinson RG, Liao F, Clegg LX, et al. (1997) Physical activity and
390 incidence of coronary heart disease in middle-aged women and men. Med Sci Sport Exerc 29:
391 901–909.

392 31. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. (2017) Genome-wide genetic data on
393 ~500,000 UK Biobank participants. bioRxiv.

394 32. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. (2016) A reference panel of
395 64,976 haplotypes for genotype imputation. Nat Genet.

396 33. UK10K Consortium (2015) The UK10K project identifies rare variants in health and disease.
397 Nature.

398 34. Howie B, Marchini J, Stephens M, Chakravarti A (2011) Genotype Imputation with Thousands of
399 Genomes. G3 GenesGenomesGenetics 1: 457–470.

400 35. Townsend P (1987) Deprivation. J Soc Policy 16: 125–146.

401 36. Smith GD, Whitley E, Dorling D, Gunnell D (2001) Area based measures of social and economic
402 circumstances: cause specific mortality patterns depend on the choice of index. J Epidemiol
403 Community Health.

404 37. Shungin D, Winkler T, Croteau-Chonka D, Ferreira T, Mägi R, et al. (2015) New genetic loci link
405 adipose and insulin biology to body fat distribution. Nature 518: 187–196.

406 38. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, et al. (2015) Second-generation
407 PLINK: rising to the challenge of larger and richer datasets. Gigascience 4: 7.

408 39. Team RDC (2011) R: A language and environment for statistical computing.

409 40. Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017) FUMA: Functional mapping and
410 annotation of genetic associations. bioRxiv.

411 41. Consortium TGte (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene
412 regulation in humans. Sci 348: 648–660.

413 42. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, et al. (2017) LD Hub: a
414 centralized database and web interface to perform LD score regression that maximizes the
415 potential of summary level GWAS data for SNP heritability and genetic correlation analysis.
416 Bioinformatics 33: 272.

417 43. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) An Atlas of Genetic
418 Correlations across Human Diseases and Traits. bioRxiv.

419 44. Pericak-Vance MA, Bebout JL, Gaskell Jr. PC, Yamaoka LH, Hung WY, et al. (1991) Linkage
420 studies in familial Alzheimer disease: evidence for chromosome 19 linkage. AmJHumGenet.

421 45. Saunders AM, Strittmatter WJ, Schmechel D, St. George-Hyslop PH, Pericak-Vance MA, et al.
422 (1993) Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic alzheimer's
423 disease. Neurology.

424 46. Thompson PD, Tsongalis GJ, Seip RL, Bilbie C, Miles M, et al. (2004) Apolipoprotein e genotype
425 and changes in serum lipids and maximal oxygen uptake with exercise training. Metabolism 53:
426 193–202.

427 47. Dudbridge F, Fletcher O (2014) Gene-environment dependence creates spurious gene-
428 environment interaction. Am J Hum Genet 95: 301–307.

429 48. Smith JC, Nielson KA, Woodard JL, Seidenberg M, Durgerian S, et al. (2014) Physical activity
430 reduces hippocampal atrophy in elders at genetic risk for Alzheimer's disease. Front Aging
431 Neurosci 6.

432 49. Podewils L, Guallar E, Kuller L, Fried L, Lopez O, et al. (2005) Physical activity, APOE

433 genotype, and dementia risk: findings from the cardiovascular health cognition study. Am J
434 Epidemiol 161.

435 50. Rovio S, Kåreholt I, Helkala E-L, Viitanen M, Winblad B, et al. (2005) Leisure-time physical
436 activity at midlife and the risk of dementia and Alzheimer's disease. Lancet Neurol 4: 705–711.

437 51. Lee H, Ash GI, Angelopoulos TJ, Gordon PM, Moyna NM, et al. (2015) Obesity-Related Genetic
438 Variants and their Associations with Physical Activity. Sport Med Open 1: 34.

439 52. Richmond RC, Davey Smith G, Ness AR, den Hoed M, McMahon G, et al. (2014) Assessing
440 Causality in the Association between Child Adiposity and Physical Activity Levels: A Mendelian
441 Randomization Analysis. PLoS Med 11.

442 53. Aschard H, Vilhjalmsson BJ, Joshi AD, Price AL, Kraft P (2015) Adjusting for heritable
443 covariates can bias effect estimates in genome-wide association studies. Am J Hum Genet 96:
444 329–339.

445 54. Day FR, Loh P-R, Scott RA, Ong KK, Perry JR (2016) A Robust Example of Collider Bias in a
446 Genetic Association Study. Am J Hum Genet 98: 392–393.

447 55. Yaghootkar H, Bancks MP, Jones SE, McDaid A, Beaumont R, et al. (2017) Quantifying the
448 extent to which index event biases influence large genetic association studies. Hum Mol Genet 26:
449 1018–1030.

450

451

452 **Figure Legends**

453 **Figure 1:** Manhattan plot of GWAS for self-reported MVPA and VPA and for accelerometer-based
454 average accelerations and fraction of accelerations > 425 mg. Negative log10-transformed p-value for
455 each SNP is plotted by chromosome and position (x-axis). The red and blue horizontal lines represent
456 thresholds for genome-wide significant and suggestive associations, respectively.

457

458 **Figure 2:** Genetic correlation of PA variables with other traits and diseases. Traits/diseases are ordered
459 from top to bottom in order of increasing p-value for the ten traits/diseases with strongest degree
460 (according to p-value) of genetic correlation with respective measure of PA. Horizontal position of bars
461 corresponds to the genetic correlation (r_g) between PA and the respective trait/disease. Error bars
462 represent 95% confidence intervals for r_g estimates. Bright green bars represent traits that showed a
463 correlation with p-value <2.5E-4, and light green bars represent traits with genetic correlation p<0.05. We
464 excluded highly redundant traits (e.g. obesity, overweight) after leaving higher ranked one in.

465

466 **Figure 3:** Results of tissue enrichment analysis using eQTL results from GTEx RNA-seq data for PA-
467 associated loci.

468 **Table 1: Summary of PA phenotypes in the UK Biobank**

Self-Report	
MVPA	Mean=1,646; SD=2,075; n=277,656
Vigorous PA: 0 vs. ≥3 days/week	71,739 cases; 119,866 controls
Accelerometry	
Average acceleration (milli-gravities)	Mean=27.93; SD=8.15; n=67,808
Fraction accelerations > 425 milli-gravities	Mean=0.0026 ; SD=0.0033 ; n=67,565

469 SD: standard deviation

470

471

472 **Table 2:** Summary of polymorphisms identified in the UK Biobank with association statistics in ARIC.

SNP ID	Chr.	Gene/Nearest Gene	Position	EA	EAF	Beta/OR	p-value	n	ARIC replication	
									Beta	p-value
MVPA										
rs429358	19	*APOE*	45,411,941	T	0.84	-0.023	4.98E-11	277,656	-0.026	0.260
rs72737787	9	*ZCCHC7*	37,336,741	G	0.65	-0.017	3.66E-10	277,656	-0.026	0.049
rs1043595	7	*CALU*	128,410,012	G	0.72	0.017	1.48E-09	277,656	0.014	0.310
rs1921981	21	*BACE2*	42,422,547	G	0.67	0.016	4.03E-09	277,656	0.021	0.106
rs3129981	6	HCG20	30,758,857	C	0.83	0.020	5.13E-09	277,656	0.007	0.730
rs921917	7	*C7orf72*	50,228,738	C	0.51	-0.015	1.19E-08	277,656	-0.007	0.568
rs12147808	14	*C14orf64*	98,618,705	C	0.79	-0.018	1.30E-08	277,656	-0.009	0.484
rs124672	8	*KCNK9*	140,722,178	A	0.71	-0.016	2.22E-08	277,656	0.003	0.840
rs148854222	5	*SLC1A3*	36,505,944	T	0.997	-0.137	2.87E-08	277,656	0.066	0.645
rs12142550	1	*HAX1*	154,258,549	C	0.75	0.016	4.60E-08	277,656	-0.011	0.461
Vigorous PA: 0 vs. ≥3 days/week										
rs6955240	7	*EXOC4*	133,581,873	G	0.61	1.044	9.57E-10	191,473	0.001	0.905
rs6667222	1	*HAX1*	154,253,661	A	0.75	1.046	9.55E-09	191,473	-0.014	0.340
rs6909774	6	*MMS22L*	97,687,471	A	0.25	1.044	3.34E-08	191,473	-0.035	0.014
Accelerometry – Average acceleration										
rs62055545	17	*MAPT-AS1*	43,964,561	C	0.78	-0.038	1.41E-09	67,808	0.006	0.680
rs34517439	1	*FUBP1*	78,450,517	C	0.88	0.044	4.23E-08	67,808	-0.010	0.635
rs4747438	10	*DNAJC1*	22,124,263	C	0.32	-0.030	4.62E-08	67,808	0.009	0.485
Accelerometry – Fraction accelerations > 425 milli-gravities										
rs10851869	15	*PML*	74,331,083	T	0.57	0.027	3.66E-08	67,565	0.015	0.227

473 EA refers to effect allele that Beta/OR corresponds to. EAF: effect allele frequency; proxy rs2732705 SNP was used
474 for rs62055545 in ARIC
475

476 **Table 3:** Loci consistently associated with PA across all four measures (each p<0.001).

					MVPA		VPA		AA		AF>425		ARIC PA	
SNP	Chr.	Nearest Gene	Position	Effect Allele	Beta	p	OR	p	Beta	p	Beta	p	Beta	p
rs7910002	10	*DNAJC1*	22050570	G	-0.01	8.7E-04	0.97	2.8E-04	-0.03	7.9E-08	-0.02	1.0E-04	0.005	0.716
rs10135643	14	*DCAF5*	69517406	A	0.01	6.5E-04	1.03	1.7E-04	0.02	4.6E-04	0.02	2.2E-04	0.023	0.064
rs5742915	15	*PML*	74336633	T	-0.01	3.7E-06	0.96	1.3E-07	-0.02	2.4E-05	-0.02	3.0E-06	0.002	0.873

477 AA: Average acceleration, AF>425: acceleration fraction greater than 425mg, ARIC PA: Physical
478 activity in ARIC study.

479

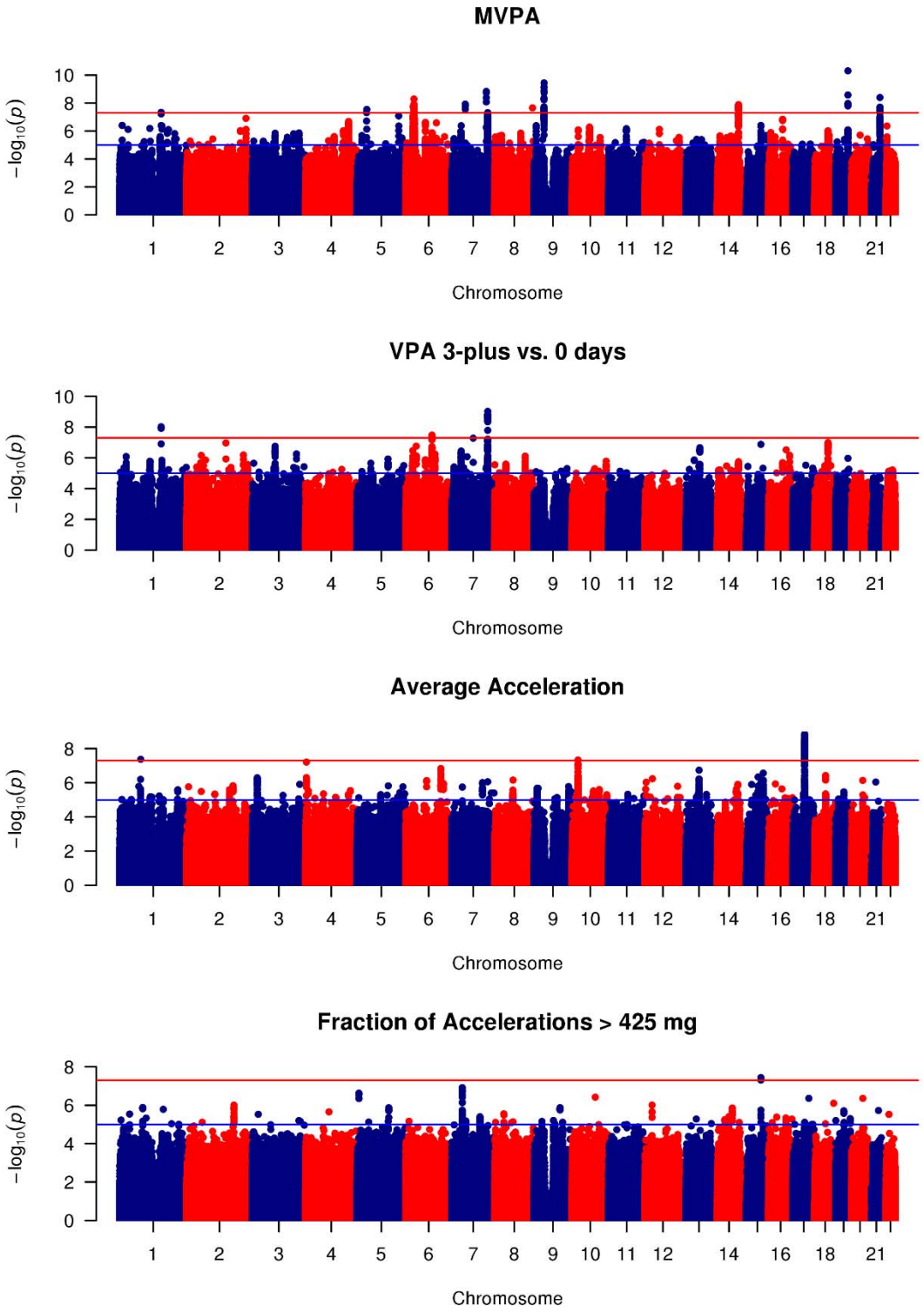
480 **Table 4:** Loci consistently associated with all four PA variables and with waist circumference from
481 GIANT consortium (each p<0.01)
482

					MVPA		VPA		AA		AF>425		Waist circum.	
SNP	Chr	Nearest Gene	Position	Effect Allele	Beta	p	OR	p	Beta	p	Beta	p	Beta	p
rs12993490	2	*CPS1*, *ERBB4*	211,953,629	C	0.01	4.1E-03	1.03	3.1E-04	0.02	4.4E-04	0.01	9.1E-03	-0.016	1.1E-03
rs7910002	10	*DNAJC1*	22,050,570	G	-0.01	8.7E-04	0.97	2.8E-04	-0.03	7.9E-08	-0.02	1.0E-04	0.015	7.5E-04
rs7250389	19	*CCNE1*	30,325,332	G	0.01	2.4E-03	1.03	1.7E-04	0.03	3.7E-06	0.02	3.5E-05	-0.02	2.1E-05

483 AA: Average acceleration, AF>425: acceleration fraction greater than 425mg

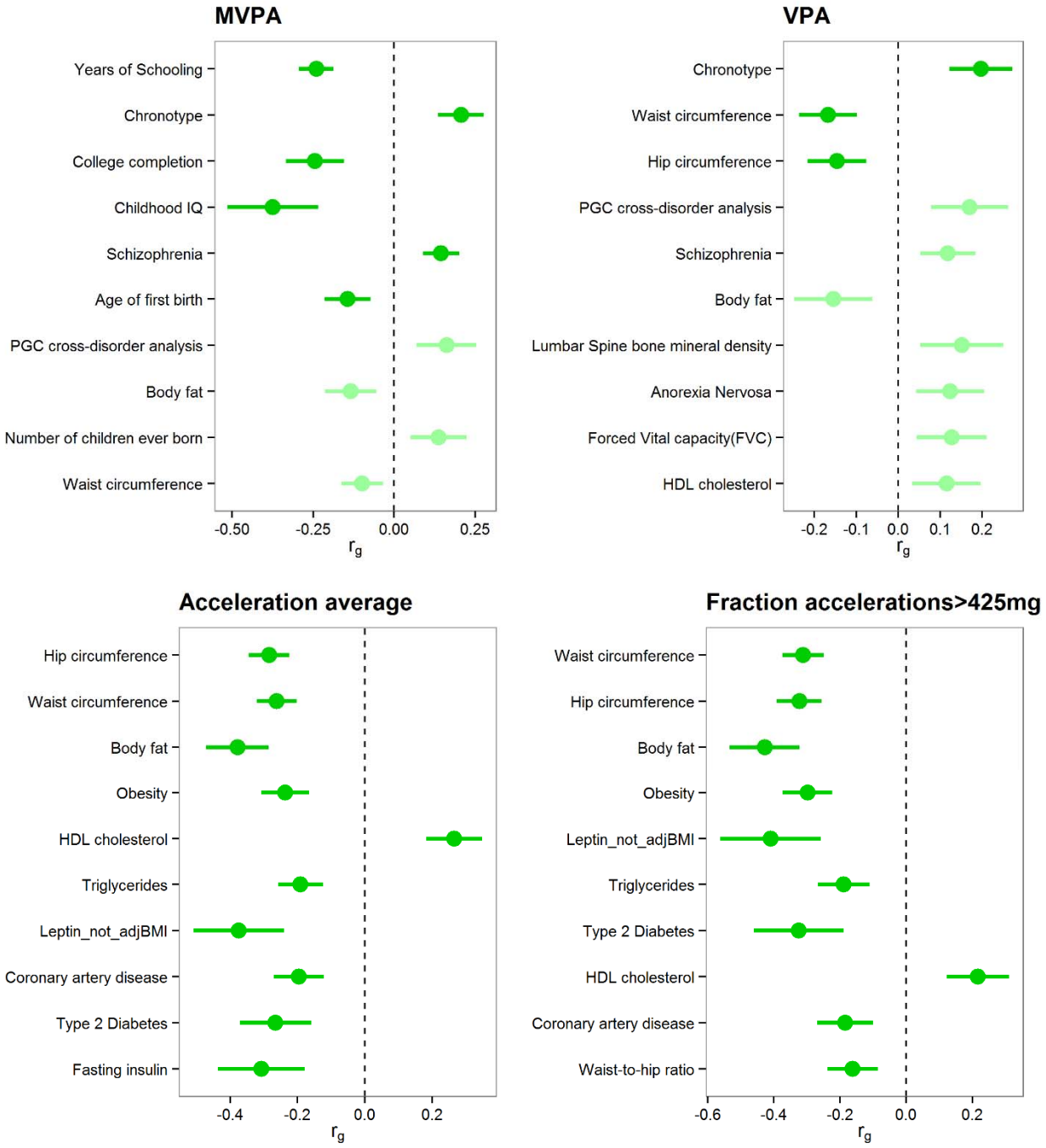
484
485

486 **Figure 1**



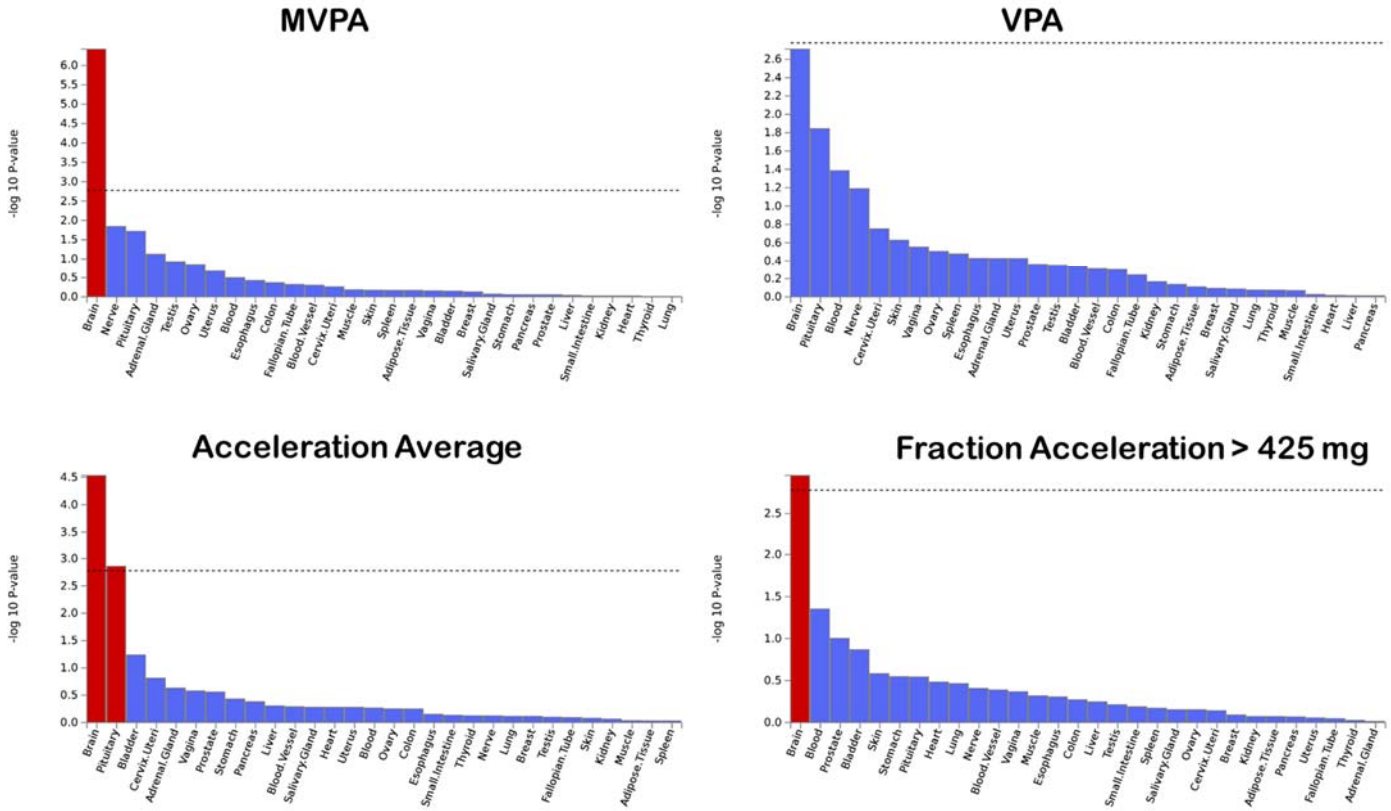
487

488 **Figure 2**



489

490 **Figure 3**



491