Genome-wide association studies of 27 accelerometry-derived physical activity measurements identifies novel loci and genetic mechanisms

Guanghao Qi,[1]† Diptavo Dutta,[1]† Andrew Leroux,[1,2] Debashree Ray,[1,3] John Muschelli,[1] Ciprian Crainiceanu[1] and Nilanjan Chatterjee[1,4]*

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[2]Department of Biostatistics and Informatics, University of Colorado, Aurora, CO 80045, USA

[3]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

[4]Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

†These authors contributed equally to this work

***Correspondence to Nilanjan Chatterjee (nilanjan@jhu.edu)**

Abstract

Physical activity (PA) is an important risk factor for a wide range of diseases. Previous genome-wide association studies (GWAS), based on self-reported data or a small number of phenotypes derived from accelerometry, have identified a limited number of genetic loci associated with habitual PA and provided evidence for involvement of central nervous system in mediating genetic effects. In this study, we derived 27 PA phenotypes from wrist accelerometry data obtained from 93,745 UK Biobank study participants. Single-variant association analysis based on mixed-effects models and transcriptome-wide association studies (TWAS) together identified 6 novel loci that were not detected by previous studies. For both novel and previously known loci, we discovered associations with novel phenotypes including active-to-sedentary transition probability, light-intensity PA, activity during different times of the day and proxy phenotypes to sleep and circadian patterns. Follow-up studies indicated the role of the blood and immune system in modulating the genetic effects and a secondary role of the digestive and endocrine systems.

38 **Introduction**

39

40 Regular physical activity (PA) is associated with lower risk of a wide range of diseases, including
41 cancer, diabetes, cardiovascular disease[1], Alzheimer's disease[2], as well as mortality[3,4]. However,
42 studies have indicated that large majority of US adults and adolescents are insufficiently active[5],
43 and thus PA interventions have great potential to improve public health. PA was shown to have
44 a substantial genetic component, and understanding its genetic mechanism can inform the
45 design of individualized interventions[6,7]. For example, people who are genetically pre-disposed
46 to low PA may benefit more from early and more frequent guidance.

47

48 A number of previous genome-wide association studies (GWAS) on physical activity have relied
49 on self-reported phenotypes, which are subject to perception and recall error[8–11]. Recently,
50 wearable devices have been used extensively to collect physical activity data objectively and
51 continuously for multiple days. To date, there have been two GWAS based on acceleromtery-
52 derived activity phenotypes. Both studies used data from the UK Biobank study[12,13] but only
53 focused on a few summaries of these high-density PA measurements. One study considered
54 two accelerometry-derived phenotypes (average acceleration and fraction accelerations > 425
55 milli-gravities) and identified 3 loci associated with PA[11]. A second study used a machine
56 learning approach to extract PA phenotypes, including overall activity, sleep duration,
57 sedentary time, walking and moderate intensity activity[14]. This study identified 14 loci
58 associated with PA and found that the central nervous system (CNS) plays an essential role in

59 modulating the genetic effects on PA. However, both studies used a small number of
60 phenotypes, which may not capture the complexity of PA patterns.
61
62 Recent studies suggest that in addition to the total volume of activity, other PA summaries may
63 be strongly associated with human health and mortality risk. For example, the transition
64 between active and sedentary states was strongly associated with measures of health and
65 mortality[4,15]. PA relative amplitude, a proxy for sleep quality and circadian rhythm, was strongly
66 associated with mental health[16]. Moderate-to-vigorous PA (MVPA) and light intensity PA (LIPA)
67 have also been reported to be associated with health[17,18]. Thus, there is increasing evidence
68 that objectively measured PA in the free-living environment is a highly complex phenotype that
69 requires a large number of summaries that provide complementary information. Understanding
70 the genetic mechanisms behind these summaries is critical for understanding the genetic
71 regulation of activity behavior and informing targeted interventions.
72
73 In this paper, we conducted genome-wide association analysis using 27 accelerometry-derived
74 PA measurements from UK Biobank data[12,13]. The phenotypes cover a wide range of features
75 including volumes of activity, activity during different times of the day, active to sedentary
76 transition probabilities and various principal components (**Supplementary Table 1**). We
77 conducted GWAS using a mixed-model-based method, fastGWA[19], to identify variants
78 associated with the above phenotypes. We also conducted transcriptome-wide association
79 studies (TWAS)[20] across 48 tissues to identify genes and tissues harboring the associations. We
80 further conducted tissue-specific heritability enrichment[21,22], gene-set enrichment[23] and

genetic correlation[24] analyses to further reveal the underlying biological mechanisms. We identified 6 novel loci associated with PA and showed that, in addition to the CNS, blood and immune related mechanisms could play an important role in modulating the genetic effects on activity, and digestive and endocrine tissues could play a secondary role.

Results

Genetic Loci Associated with Physical Activity

Single-variant genome-wide association analysis identified a total of 16 independent loci, including five novel ones compared to previous studies (**Table 1 and Fig. 1**). The locus indexed by single-nucleotide polymorphism (SNP) rs301799 on chromosome 1 was associated with total log acceleration (TLA) between 6pm to 8pm, representing early evening activity. Three novel loci were discovered on chromosome 3: the locus indexed by rs3836464 was associated with active-to-sedentary transition probability (ASTP); the locus indexed by rs9818758 was associated with relative amplitude, which is a proxy sleep behavior and circadian rhythm[25]; the locus indexed by insertion-deletion (INDEL) 3:131647162_TA_T (no rsid available) was associated with TLA 2am-4am which is a proxy phenotype for activity during sleep. LIPA appeared to be associated with other SNPs near 3:131647162_TA_T but not the lead variant itself, indicating multiple independent signals at the same locus (**Fig. 1**). Another locus indexed by rs2138543 is associated with TLA 6am-8am which represents early morning activity, and the

second principal component of log-acceleration (PC2), for which the interpretation is less clear (**Table 1**).

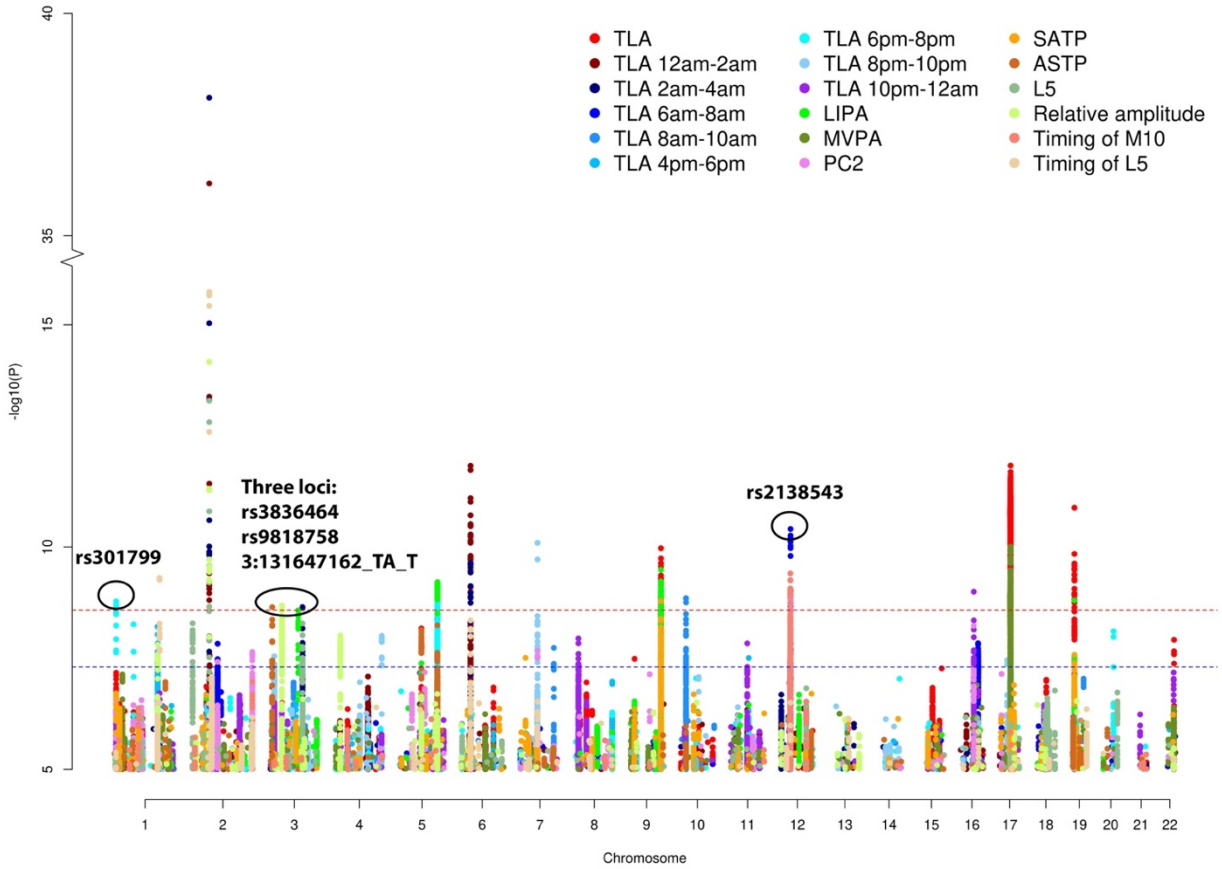


Fig. 1 Manhattan plot for 18 traits that are significantly associated with at least one variant at $p < 2.63 \times 10^{-9}$ in single-variant analysis. The red dashed line is $p = 2.63 \times 10^{-9}$ accounting for the number of independent traits. The blue dashed line is the standard genome-wide significance threshold $p = 5 \times 10^{-8}$. Five novel loci that have not been discovered in previous GWAS of physical activity, sleep duration and circadian rhythm are circled out (see Table 1).

Table 1. Significant loci associated with physical activity in single-variant analysis.

Lead variant[a]	Chromosome region	Base pair	Minor allele	Major allele	Minor allele frequency	Significantly associated traits and p-values[c]	Previous studies that discovered the locus
Novel loci[b]							
rs301799	1p36.23	8489302	C	T	0.422	TLA 6pm-8pm (1.7e-09)	NA
rs3836464	3p25.3	10454772	CA	C	0.274	ASTP (2.2e-09)	
rs9818758	3p21.31	49382925	A	G	0.171	Relative amplitude (2.1e-09)	
3:131647162_TA_T	3q22	131647162	T	TA	0.466	TLA 2am-4am (2.2e-09)	
rs2138543	12q12	39298423	A	T	0.477	TLA 6am-8am (3.9e-11), PC2 (1.5e-09)	
Known loci							
rs1144566	1q25.3	182569626	T	C	0.030	Timing of L5 (5e-10)*	4
rs113851554	2p14	66750564	T	G	0.050	TLA 12am-2am (6.7e-37)*, TLA 2am-4am (7.9e-39)*, L5 (1.3e-33)*, Relative amplitude (6.9e-15)*, Timing of L5 (5.4e-22)*	1,4
rs2909950	5q33.1	151886147	A	G	0.418	TLA 6pm-8pm (9.4e-10)*	4
rs12717867	5q33.1	152412845	G	A	0.453	LIPA (6.2e-10)*	4
rs9369062	6p21.2	38437303	C	A	0.292	TLA 12am-2am (1.5e-12)*, TLA 2am-4am (2.3e-10)*	4
rs2006810	7q11.22	69902152	C	T	0.395	TLA 8pm-10pm (8.1e-11)*	1,4
rs1268539	9q33.3	128195657	A	C	0.419	TLA (1.1e-10), LIPA (3.2e-10)*	1
rs564819152	10p12.31	21820650	G	A	0.320	TLA 8am-10am (1.4e-09)*	1,3
rs12927162	16q12.2	52684916	G	A	0.277	TLA 10pm-12am (1e-09)*	4
rs2532402	17q21.31	44304130	G	C	0.221	TLA (1.5e-12), MVPA (1.9e-10)	1,2,3,4
rs3837946	19p13.2	9955920	TTTTG	T	0.475	TLA (1.3e-11), LIPA (1.6e-09)*	1,3

[a]Significant variants (single nucleotide polymorphism (SNP) or insertion/deletion) are defined as those with fastGWA p-value $< 2.63 \times 10^{-9}$ (Bonferroni corrected for 19 independent traits). LD clumping was performed at $r^2 < 0.1$ and lead variants of different loci were required to be >500kb apart.

[b]A locus is defined as novel if its lead variant is >500kb from the lead variant of any loci identified in one of the previous GWAS: 1) Doherty et al study on a smaller set of accelerometry-based physical activity[14]; 2) Klimentidis et al study on self-reported and accelerometry based physical activity[11]; 3) Dashti et al study on self-reported sleep duration[30]; 4) Jones et al study on circadian rhythm[26]. Loci that were discovered by the four above studies are marked with 1, 2, 3 and 4 in the last column, respectively. Novel phenotypes associated with known loci are marked with *.

[c]TLA: total log-acceleration. ASTP: active-to-sedentary transition probability. MVPA: moderate-to-vigorous physical activity. PC2: second principal component. LIPA: light-intensity physical activity.

120 Our analysis also identified novel phenotypes for several known loci (**Table 1**). The strongest
121 signal was seen for the locus indexed by rs113851554 which is associated with multiple sleep
122 and circadian rhythm proxy phenotypes including TLA 12am-2am ($p = 6.7 \times 10^{-37}$), TLA 2am-
123 4am ($p = 7.9 \times 10^{-39}$), average log acceleration during the least active 5 hours of the day (L5,
124 $p = 1.3 \times 10^{-33}$), timing of L5 ($p = 5.4 \times 10^{-22}$) and PA relative amplitude ($p = 6.9 \times 10^{-15}$).
125 This locus was previously identified to be associated with accelerometry-derived sleep duration
126 in UK Biobank[14]. Among other known loci, 5 were only discovered in the GWAS of self-reported
127 circadian rhythm[26] but not in the other studies considered (**Table 1**, last column). In our
128 analysis, the loci indexed by rs1144566, rs9369062 and rs12927162 were associated with sleep
129 proxy phenotypes including timing of L5, TLA 12am-2am, TLA 2am-4am and TLA 10pm-12am.
130 Two other loci, indexed by rs2909950 and rs12717867, were associated with TLA 6pm-8pm and
131 LIPA, respectively.
132
133 *Transcriptome-Wide Association Study and Colocalization Analysis*
134
135 We performed transcriptome-wide association studies (TWAS) [20,27] for each PA trait based on
136 gene expression data across 48 tissues available through GTEx (version 7)[28]. Our analysis
137 identified 15 loci (**Table 2, Fig. 2, Supplementary Table 2**) with significant association in at least
138 one trait-tissue pair analysis after correcting for multiple testing (Benjamini-Hochberg corrected
139 p-value < 2.5×10^{-6}, see **Methods**). We identified a novel locus and an underlying index
140 pseudogene *PDXDC2P (16q22.1)*, the expression of which in esophagus mucosa and EBV
141 transformed lymphocytes appeared to be genetically associated with TLA 6am-8am (**Fig. 2**). The

142 locus was not previously reported by any prior GWAS and were not close to any of the 5 novel
143 regions detected by our single variants analysis (**Table 2**).
144
145 The TWAS analysis also identified novel PA phenotypes, potential target genes and underlying
146 tissues for many of the known loci or novel loci detected through single variant analysis (**Table**
147 **2**). Consistent with a previous study[14], the TWAS analysis showed that genetic association for
148 PA traits often points towards involvement of CNS (**Table 2**). Further, our analysis indicates
149 consistent involvement of blood and immune, digestive and endocrine systems in modulating
150 the genetic effects on PA. Among the 15 loci significant in TWAS analysis, the lead genes of 4
151 loci were significantly associated with PA phenotypes via the blood and immune tissues. For
152 example, the genetically predicted expression of *PBX3* and *KANSL1* in the blood and immune
153 tissues were each associated with 3 PA phenotypes. The genes associated with PA via blood and
154 immune tissues were also associated via digestive (e.g., esophagus mucosa) and/or endocrine
155 (e.g., thyroid) tissues, but only one of them overlapped with the 9 genes that were associated
156 with PA phenotypes through the CNS (**Table 2**). Another locus, represented by *C3orf62*, were
157 associated with PA relative amplitude only via the digestive and endocrine tissues but not the
158 blood/immune tissues or CNS. These findings suggested that the genetic regulation of PA
159 occurs via at least two different pathways: a primary pathway involving the CNS (brain in
160 particular) and a secondary pathway involving the blood/immune system and, potentially, the
161 digestive and endocrine systems.
162
163

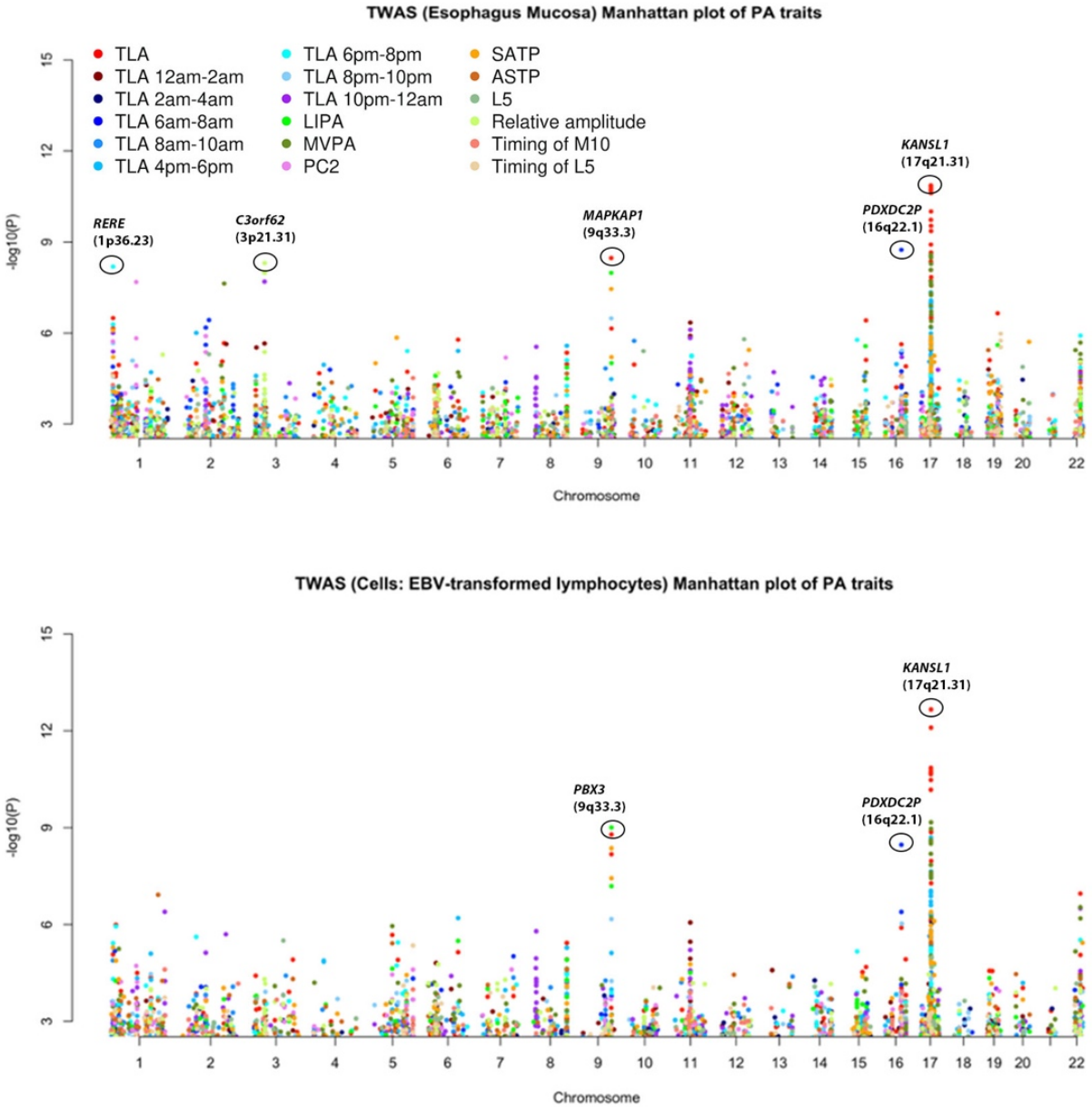


Fig. 2 TWAS Manhattan plots for two tissues that harbor the novel TWAS locus represented by *PDXDC2P*. The figure shows -log10 p-value for all the genes expressed in esophagus mucosa (digestive, upper panel) or EBV-transformed lymphocyte cells (blood/Immune, lower panel) as reported in GTEx v7. Significant genes with FDR corrected p-values$< 2.5 \times 10^{-6}$ are circled (see Methods for details). Only the PA traits that are significantly associated with at least one variant at $p < 2.63 \times 10^{-9}$ in single-variant analysis are shown.

Table 2. Significant loci associated with physical activity identified in TWAS.

Most significant gene[a]	CHR	Gene start	Gene end	Locus	Minimum p-value across traits and tissues	Secondary genes in the locus with comparable associations[c]	Trait and tissue[d]	Previous studies that discovered the locus
Novel loci[b]								
PDXDC2P	16	70069541	70098679	16q22.1	1.81E-09	-	TLA 6am-8am (Blood/Immune, Digestive)	NA
Known loci								
RERE†	1	8412457	8877702	1p36.23	5.28E-09	*ENO1-IT1, RP5-1115A15.1*†	TLA 6pm-8pm (Digestive, Other, Blood/Immune)	4,GWAS
RN7SKP16	1	33802167	33802465	1p35.1	2.72E-09	-	MVPA (CNS)	3
RP5-1160K1.6	1	110171118	110171939	1p13.3	7.68E-10	-	PC2 (CNS)	4
C3orf62†	3	49306219	49315263	3p21.31	3.28E-10	*RP11-694I15.7, ARIH2, DAG1, SLC25A20*	Relative amplitude (Digestive, Endocrine)	4,GWAS
CTC-467M3.3	5	87988462	87989789	5q14.3	1.39E-08	-	ASTP (CNS)	1,4
NMUR2	5	151771093	151812929	5q33.1	8.88E-10	-	TLA 6pm-8pm (CNS, Other)	4,GWAS
PBX3†	9	128509624	128729656	9q33.3	3.04E-10	*MAPKAP1*	LIPA (Blood/Immune, Digestive, Endocrine, Other); SATP (Blood/Immune, Other, Digestive, Endocrine); TLA (Blood/Immune, Other, Digestive, Endocrine)	1,GWAS
DNAJC1	10	22045466	22292698	10p12.31	3.20E-09	*CASC10*†	TLA 8am-10am (CNS)	1,3,GWAS
RP11-396F22.1†	12	39300253	39303394	12q12	9.89E-10	-	Timing of M10 (CNS), PC2 (CNS, Other), TLA 6am-8am (CNS, Other)	3,4,GWAS
FMNL1	17	43299590	43324633	17q21.31	2.82E-09	*CTD-2020K17.1*	TLA (CNS)	1,2,3,4
KANSL1	17	44107282	44302733	17q21.31	2.17E-13	*KANSL1-AS1, RP11-798G7.8*	MVPA (CNS, Blood/Immune, Other, Digestive, Musculoskeletal/connective); TLA 4pm-6pm (Blood/Immune, Other, Digestive); TLA (Adipose, CNS, Blood/Immune, Other, Digestive, Cardiovascular, Musculoskeletal/connective, Endocrine)	1,2,3,4,GWAS
ZNF846	19	9862669	9903856	19p13.2	2.22E-10	*OLFM2, PIN1, CTD-2623N2.3*	TLA (Other)	1,3,4,GWAS
JUND	19	18390563	18392432	19p13.11	3.18E-09	-	SATP (Other)	4
LINC00634	22	42348169	42354937	22q13.2	4.85E-11	-	TLA (CNS)	4

[a]For each gene, we apply Benjamini-Hochberg (BH) correction across all trait-tissue pairs. Significant genes are defined as those with BH corrected p-value $< 2.5 \times 10^{-6}$. Significant genes are clustered using an approach similar to LD clumping so that different loci, marked by the transcription start site (TSS) of the lead gene, are >1Mb apart. Gene symbols are italicized. † Genes of which the eQTL signal colocalizes with GWAS signal of the most significantly associated phenotype.

[b]A locus is defined as novel if its cis region ($\pm$1Mb from TSS) does not harbor a variant reported in Table 1 or any of the four previous studies considered in Table 1. Loci that were discovered by the above studies are marked with 1, 2, 3 and 4 in the last column, respectively. Loci that were discovered by our single-variant analysis (Table 1) are marked with "GWAS" in the first column.

[c]Other genes in the cluster with minimum p-value less that 10 times the minimum p-value of the lead gene are shown in column "Secondary genes in the locus with comparable associations". See Supplementary Table 2 for all the significant tissue-trait pairs.

[d]CNS: central nervous system. TLA: total log-acceleration. ASTP: active-to-sedentary transition probability. SATP: sedentary-to-active transition probability. MVPA: moderate-to-vigorous physical activity. PC2: second principal component.

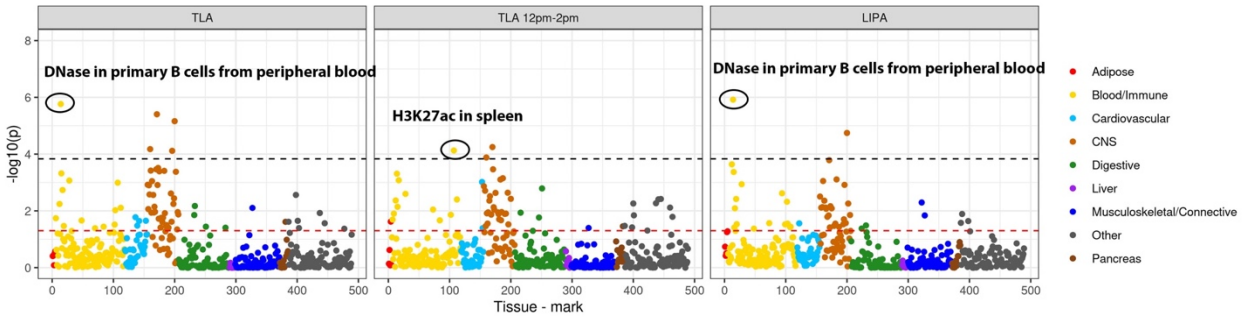
182 Several genes that were found to be significantly associated to specific PA traits in our TWAS
183 analysis, were also found to be highly overlapping with genes that were previously reported to
184 be associated with various traits and diseases including but not limited to neuropsychiatric
185 diseases, behavioral traits, anthropometric traits and autoimmune diseases (**Supplementary**
186 **Fig. 1**). For example, we found that the genes associated with TLA across different tissues, are
187 enriched for genes that have been associated with neuroticism, bipolar disorder, Parkinson's
188 disease, cognitive function and several others indicating the putative involvement of the CNS in
189 the genetic mechanism of TLA. Additionally, the genes associated with relative amplitude
190 overlapped highly with those associated with several autoimmune diseases like inflammatory
191 bowel disease, ulcerative colitis in addition to different behavioral and cognitive traits
192 (**Supplementary Fig. 1**). These results further supported the possible involvement of both CNS
193 as well as the blood and immune system in the genetic mechanism of PA traits.
194
195 We performed a colocalization analysis to gain further insights on the tissue specific activity of
196 the significant genetic loci. Among the 16 loci significantly associated with PA, 9 loci colocalized
197 with the eQTL signals for at least one gene and one tissue with a colocalization probability
198 (PP4) > 0.8 (**Supplementary Table 3**). Colocalization occurred in a similar set of tissues as those
199 that harbored the TWAS associations (**Table 2 and Supplementary Table 3**), namely the CNS,
200 blood and immune, digestive and endocrine tissues, and also in a number of cardiovascular
201 tissues that were not highlighted by TWAS. Among the 15 lead genes for TWAS significant loci,
202 the eQTL signal of 4 genes (*RERE, C3orf62, PBX3* and *RP11-396F22.1*) colocalized with PA GWAS

203 signal in at least one tissue. Colocalization also occurred in two other secondary genes (*RP5-*

204 *1115A15.1* and *CASC10*).

205

206



207

208 **Fig. 3 Tissue-specific heritability enrichment p-values for traits with significant enrichment at**

209 **FDR < 0.05 in blood and immune tissues.** The analysis was conducted using tissue/cell type

210 specific stratified LD score regression based on 6 chromatin-based annotations in 111 tissues

211 and cell types described in Finucane et al, Nature Genetics 2018 (PMID: 29632380). Each dot

212 corresponds to an annotation in a tissue or cell type. A complete list of tissue and cell types is

213 provided in Supplementary Table 7 of the above paper. Black line corresponds to FDR < 0.05 (-

214 log(p-value)=3.83) across all combinations of trait, tissue, and histone mark. Red line

215 corresponds to p = 0.05. See Supplementary Fig. 4 for the enrichment p-values for the rest of

216 the traits. CNS: central nervous system.

217

218

219 *Analysis of Heritability and Co-Heritability*

220

221 Our fastGWA analysis estimated genome-wide heritability of PA phenotypes as an intermediate
222 output. The estimates appeared to be dependent on the sparsity level of the genetic
223 relationship matrix (**Supplementary Fig. 2**). We chose the results under the lower cutoff (0.02)
224 since it captured more subtle relatedness and should give more accurate heritability estimates.
225 The estimates of heritability varied across different PA phenotypes. A number of traits were
226 estimated to have higher heritability than others, including TLA (0.15), TLA 6pm-8pm (0.15),
227 MVPA (0.14) (**Supplementary Fig. 2**). Afternoon and pre-sleep evening activity (TLA 4pm to
228 12am) appeared to be more heritable than morning activity (TLA 2am to 12pm). As could be
229 expected, phenotypes with higher heritability tend to have a higher average χ^2 statistic for
230 genetic associations, and a QQ plot which deviate further from the null line (**Supplementary**
231 **Fig. 3**).

232

233 We further used stratified LD-score regression for partitioning heritability by functional
234 annotations of genome[21,22]. Consistent with TWAS findings, this analysis also indicated possible
235 role for blood and immune system in addition to CNS for genetic regulation of PA (**Fig. 3**). In
236 particular, heritabilities for both TLA and LIPA were enriched for DNase I hypersensitivity sites
237 (DHS) in primary B cells from peripheral blood and that for TLA 12pm-2pm were enriched for
238 H3K27ac in spleen. We also found potential enrichment in other traits, though they were not
239 significant after FDR adjustment. For example, for TLA 8am-10am, MVPA, and ASTP the

240 heritability enrichment in active chromatin regions of blood/immune tissues were all close to
241 being statistically significant (**Supplementary Fig. 4**).
242
243 We further used LD score regression[24,29] to explore genetic correlation between PA phenotypes
244 and four broad groups of complex traits and diseases (**Supplementary Fig. 5 and**
245 **Supplementary Table 4**). Genetic correlations were identified (FDR < 10%) between PA
246 phenotypes and: (1) neurological, psychiatric and cognitive traits, including Alzheimer's disease
247 (AD), attention-deficit hyperactivity disorder (ADHD), depressive symptoms, intelligence, and
248 neo-conscientiousness; (2) auto-immune diseases, with the strongest correlation for multiple
249 sclerosis and weaker correlations for Crohn's disease and primary billary cirrhosis; (3) obesity-
250 related anthropometric traits and (4) cholesterol levels. These results broadly supported our
251 previous results indicating the role of CNS and blood/immune related mechanisms in the
252 genetics of PA traits.
253
254 **Discussion**
255
256 In summary, our study provided novel insights to genetic architecture of physical activity
257 through genome-wide association analysis of an extensive set of accelerometry based PA
258 phenotypes, derived in the UK biobank study, and a series of follow-up genomic analyses. We
259 identified a total of six novel loci, most of which were associated with PA phenotypes not
260 considered in previous studies [11,14,26,30]. Our analysis also identified novel phenotypes
261 associated with the known loci. Further, we provided multiple independent lines of evidence

262 that genetic mechanisms for association for PA involve the blood and immune system, which
263 can be potential targets for developing future interventions.
264
265 Compared to the 15 loci identified by the two previous GWASs on accelerometry-based PA [11,14],
266 the novel loci we discovered have increased the number of PA susceptibility loci by 40%. Most
267 of the novel loci were connected to the expression of genes, pseudogenes or long non-coding
268 RNAs (lncRNA, **Table 1 and 2**). The novel locus index by rs301799 overlapped with the TWAS
269 locus indexed by *RERE* (**Table 1 and 2**). The *RERE* gene was shown to be important for early
270 development of brain, eyes, inner ear, heart and kidneys, which could have complex effects on
271 an individual's ability to perform physical activity [31]. The novel locus indexed by rs9818758
272 overlaps with the TWAS locus index by *C3orf62*. Though it was unclear how *C3orf62* is involved
273 in PA, two secondary genes in the locus, *ARIH2* and *DAG1* (**Table 2**), appeared to be involved in
274 relevant biological processes. *ARIH2* was found to be essential for embryogenesis by regulating
275 the immune system[32]; *DAG1* was found to play a role in the regeneration of skeletal muscles[33].
276 Another two novel loci were connected to pseudogene *PDXDC2P* and lncRNA *RP11-396F22.1* of
277 which the function is less clear and may be worth future lab investigation.
278
279 The novel phenotypes in this study provided important insights into the genetic architecture of
280 PA, which may have been overlooked by previous GWASs on a small number of phenotypes.
281 The accelerometry-based study by Doherty et al identified the genetic associations with overall
282 activity, sleep duration and sedentary time[14]; the study by Klimentidis et al studied the average
283 acceleration and the duration of active states[11]. Our results found that there can be different

284 genetic architecture for PA during different times of the day, and there can be unique variants
285 that only affect certain PA patterns, like ASTP, LIPA and relative amplitude, but not others
286 (**Table 2**). The heritability and genetic correlation can also vary across different PA phenotypes
287 (**Supplementary Fig. 2 and Supplementary Fig. 5**).
288
289 TWAS and tissue-specific heritability enrichment analysis suggested that in addition to the CNS,
290 the blood and immune system could be also associated with PA. This finding was further
291 supported by colocalization, gene-set enrichment and genetic correlation analyses. A previous
292 study[14], which explored enrichment of heritability for PA traits by tissue-specific gene
293 expression patterns, identified potential modulating role of the CNS, adrenal/pancreatic and
294 skeletal muscle tissues. Our study, which used a more extended set of phenotypes and
295 chromatin-state-based annotations, confirmed previous findings and further highlighted the
296 role of the blood and immune system. Previous medical literature has established the effect of
297 PA on immune functions. A study showed that higher PA is associated with elevation of T-
298 regulatory cells and lower risk for autoimmune diseases[34]. Multiple studies showed that regular
299 moderately intense PA boost immune functions in older adults and protects against age-related
300 inflammatory disorders[35–37]. Our analysis suggested the link between PA and immune functions
301 in the genetic pathways and future studies are needed to better understand the underlying
302 mechanisms and causal directions.
303
304 In addition to the blood and immune system, TWAS and enrichment analysis also suggested
305 that the digestive system and endocrine system could be involved in modulating the genetic

306 effects on PA. A previous study found that PA has complex effects on gastroinstestinal health[38]:
307 acute strenuous activity may provoke gastrointestinal symptoms while low-intensity activity
308 could have benefits. Interestingly, three TWAS loci that were significant in digestive tissues
309 were associated with PA phenotypes that are proxies for meal-time activity: *PDXDC2P* with TLA
310 6am-8am, *RERE* with TLA 6pm-8PM and *KANSL1* with TLA 4pm-6pm (**Table 2**). It was also known
311 that multiple organs in the endocrine system produce hormones that regulate physiological
312 functions of the body, which can have complex bidirectional relationships with PA[39–43]. Our
313 TWAS analysis indicated that the genes associated with PA via the blood and immune system
314 tended to also be associated with the digestive and endocrine systems, but do not usually
315 overlap with the genes associated with the CNS. This suggests that the blood and immune,
316 digestive and endocrine systems may be involved in the same broad pathway that affects PA,
317 which is different from that of CNS.
318
319 This study has a number of limitations. Though we derived a more extensive set of PA
320 phenotypes than previous studies, information was still lost when collapsing a 7-day continuous
321 times series of wrist accelerometry into 31 PA phenotypes. The ideal approach would be to
322 conduct a GWAS utilizing all the information across the 7 days of accelerometry measurements.
323 Results could outline genetic regulation of a continuous course of PA over time. The current
324 analysis of TLA during 12 non-overlapping two-hour time intervals during the day, indicated
325 that different genetic variants may affect PA during different times of the day (**Tables 1 and 2**).
326 Another limitation is that some of the phenotypes are not directly interpretable. For example,
327 the PCs of log acceleration are less interpretable than other phenotypes, such as TLA and ASTP.

However, they do reflect important features of physical activity and warrant further investigations. A potential solution is to obtain proxy measurements that are interpretable and highly correlated with PC scores.

In conclusion, we conducted association studies on a wide range of PA phenotypes and identified 6 novel loci associated with PA. We found that in addition to the CNS, the blood and immune system may also play an important role in the genetic mechanisms of PA, and the digestive and endocrine systems could also be involved in the blood and immune pathway.

Materials and Methods

Study Cohort and Physical Activity Phenotypes

The UK Biobank study consists of ~500,000 individuals in the United Kingdom with comprehensive genotype and phenotype data[13]. We used a subset of the 103,712 individuals who were invited and agreed to participate in the accelerometry sub-study where participants wore a wrist-worn accelerometer for up to 7 days[12]. Accelerometry data from participants are available at multiple resolutions. Here, the individual-specific set of accelerometry-based phenotypes was derived from the five-second level acceleration data provided by the UK Biobank team. Individuals were screened for poor quality data using indictors provided by the UK Biobank. In addition, we required individuals to have at least 3 days (12am-12am) of sufficient wear time defined as estimated wear time greater than 95% of the day (>= 1368

minutes). Our inclusion criteria for this analysis closely mirrors that described in a related paper from our group[44] with the exception that we did not exclude participants younger than 50 at the time of accelerometer wear or based on missing demographic and lifestyle data and instead excluded individuals based on ancestry and genotype data (see subsection *Genotype Data* below).

Physical activity phenotypes were all calculated at the day level and then averaged within study participants across days to obtain one measure for each phenotype and study participant. This led to in 31 PA phenotypes for 93,745 study participants that covered a wide spectrum of information: 1) total volume of activity (total acceleration (TA), total log acceleration (TLA)); 2) activity during 12 disjoint two-hour windows of the day (TLA 12am-2am, TLA 2am-4am, …, TLA 10pm-12am); 3) duration of sedentary state (ST), LIPA and MVPA; 4) PA principal components (PC1-6); 5) active-to-sedentary transition probability (ASTP) and sedentary-to-active transition probability (SATP); 6) proxy phenotypes for circadian patterns, including dynamic activity ratio estimate (DARE), activity during the most active 10 hours (M10) and least active 5 hours (L5) of the day, timing of M10 and L5, and PA relative amplitude. They included most of the phenotypes used in the previous PA association studies as well (See **Supplementary Table 1** for details)[11,14]. The exact procedure for deriving study participant-specific phenotypes is described in detail in the supplemental material of the related paper from our group[44]. The phenotypes were inverse-normal transformed

$$INT(y_i) = \Phi^{-1}\left[\frac{rank(y_i)-k}{n-2k+1}\right], \quad k = 3/8,$$

where the transformed variables have mean 0 and variance 1.

372

373 *Removing Highly Correlated Phenotypes*

374

375 Some of the initial 31 PA phenotypes were highly correlated (**Supplementary Fig. 6**). To avoid
376 counting similar phenotypes multiple times, if two phenotypes had correlation > 0.8 we
377 removed one of them. First, we removed total acceleration (TA), duration of sedentary state
378 (ST), PC1 and M10 due to their high correlation with total log acceleration (TLA). TLA was
379 retained as the main metric for the total volume of activity. Most previous studies used TA
380 instead of TLA as the main metric for the volume of activity. However, the distribution of TA is
381 highly skewed, which could lead to lower power for association testing and instability of
382 findings due to dependence on extreme observations. In total, 4 phenotypes were removed
383 and 27 PA phenotypes were retained for the association analysis.

384

385 *Genotype Data*

386

387 The imputed genotype data for ~93 million variants, using UK10K, 1000 Genomes (Phase 3) and
388 Haplotype Reference Consortium as reference panel, provided by UK Biobank were used and
389 merged with the PA phenotype data. We excluded study participants according to the following
390 criteria: 1) non-white ancestry; 2) putative sex chromosome aneuploidy; 3) an excessive
391 number of relatives (more than 10 putative third-degree relatives in the kinship table); 4)
392 sample was not in the input for phasing of chr1-chr22. After applying these exclusion criteria
393 the sample was further reduced to 88,411 study particiants for downstream analysis.

394

395 We conducted variant quality control to ensure that genetic variants with poor genotyping
396 quality do not affect the results. Specifically, variants that satisfy any of the following criteria
397 were removed: 1) imputation INFO score < 0.8; 2) MAF < 0.01; 3) Hardy-Weinberg Equilibrium
398 (HWE) p-value < 1×10^{-6}; 4) missing in more than 10% study participants. After the filtering,
399 8,951,705 variants remained for downstream analysis, of which 8,067,228 (90.1%) were single
400 nucleiotide polymorphisms (SNPs) and the rest (9.9%) were insertion-deletions (INDELs).

401

402 *Association Analysis*

403

404 We used a fast mixed-effects model method, fastGWA[19], for genome-wide association analysis.
405 Like other mixed-effects model methods, fastGWA allows the inclusion of related and unrelated
406 individuals but improves computational efficiency by incorporating a sparse genetic relationship
407 matrix (GRM). The GRM measures the genetic similarity between individuals and each element
408 is the correlation of genotypes between a pair of individuals. We constructed the GRM using
409 LD-pruned variants that had MAF > 5% and were present in HapMap3 (LD-pruning was done in
410 PLINK using the following set up as recommended in Jiang et al [19]: window size = 1000Kb, step-
411 size = 100 and r^2 = 0.9). We further computed a sparse-GRM at sparsity level 0.05 to capture
412 the genetic relatedness between the closely related individuals only and reduced others to
413 zero. We used the Haseman-Elston regression to estimate the variance of the random effects as
414 an intermediate step of fastGWA. This approach is orders of magnitude faster than the previous
415 state-of-the-art, BOLT-LMM[45,46].

416

417 Models were adjusted for age, sex and the first 20 genetic principal components as covariates.
418 Because the PA phenotypes are correlated, principal component analysis (PCA) was conducted
419 on the phenotypes to estimate the number of independent phenotypes before setting the
420 GWAS significance threshold. At least 19 phenotype PCs were needed to explain 99% percent of
421 the PA phenotypic variance (**Supplementary Fig. 7**). Variants with p-value below the threshold
422 $5 \times 10^{-8}/19 = 2.63 \times 10^{-9}$ were declared to be statistically significant, which accounted for
423 the number of independent phenotypes. LD clumping was conducted based on the minimum p-
424 value across phenotypes. The requirements for the lead SNPs of different loci were to have
425 $r^2 < 0.1$ and be at least > 500kb apart.

426

427 A locus was defined as novel if its lead variant is > 500kb from the lead variant of any known
428 loci discovered by the following GWASs on PA, sleep, and circadian rhythm: (1) Doherty et al
429 study on a smaller set of accelerometry-derived PA phenotypes[14]; (2) Klimentidis et al study on
430 self-reported and accelerometry-derived PA[11]; (3) Dashti et al study on self-reported sleep
431 duration[30]; (4) Jones et al study on circadian rhythm[26].

432

433 Transcriptome-wide association studies (TWAS) were conducted using the FUSION R program[20]
434 with reference models generated from 48 tissues of GTEx v7[28]. TWAS analysis was limited to
435 the 18 traits with at least one genome-wide significant variant ($p < 2.63 \times 10^{-9}$). Multiple
436 testing due to the large number of tissue-trait combinations (48*18=864) was addressed by a
437 two-stage adjustment approach: 1) for each variant, the Benjamini-Hochberg (BH) adjustment

438 was applied across all tissue-trait pairs; 2) each variant with BH-adjusted p-value 2.5×10^{-6}
439 was then identified (accounting for 20,000 protein-coding genes). Since there can be multiple
440 genes in close proximity to each other, to identify independent loci detected by TWAS analysis,
441 genes were clustered based on significant associations. A clumping approach was used, which
442 selected the gene with the smallest minimum p-value across tissue-trait pairs and removed the
443 other genes with a transcription start site (TSS) within 1Mb of the lead gene TSS. The process
444 continued by identifying the gene with the next smallest minimum p-value and iterating. The
445 only exception was when the lead gene of the cluster was not a protein-coding gene (e.g.,
446 pseudogene, lncRNA) and a protein-coding gene was in the cluster. In this case the protein-
447 coding gene with the smallest minimum p-value was identified as the lead gene. This led to
448 independent gene clusters at genomic loci which were least 1Mb apart, i.e., none of the lead
449 gene TSS is within the cis region of another lead gene.
450
451 *Enrichment Analysis*
452
453 Stratified LD score regression[21,22] was used to identify the tissues and genomic annotations
454 enriched by the heritability for PA. For tissue specific analysis, chromatin-based annotations
455 were used as derived from the ENCODE and Roadmap data[47,48] by Finucane et al[21]. The
456 annotations were based on narrow peaks of DNase I hypersensitivity site (DHS) and five
457 activating histone marks (H3K27ac, H3K4me3, H3K4me1, H3K9ac and H3K36me3) observed for
458 111 tissues or cell types, resulting in a total of 489 annotations. Stratified LD score regression

459 computes the heritability attributed to each annotation and computes a coefficient and a p-
460 value that characterize enrichment.
461
462 In a separate analysis, the enrichment of TWAS signals was evaluated among the genes that
463 have been reported to be associated to different traits, using FUMA[23,49]. For a given PA trait, we
464 defined a gene-set as the genes that were significant at an exome-wide level ($p < 2.5 \times 10^{-6}$)
465 and investigated whether these genes overlapped with the genes that have been mapped to
466 genome-wide significant variants for different traits as reported in GWAS catalog[50]. The
467 collection of such genes have been detailed in Molecular Signatures Database (MSigDB)[51]. We
468 used FUMA to compute the proportion of genes related to other diseases and traits that were
469 also identified by our TWAS analysis and computed enrichment p-values using the Fisher's exact
470 test.
471
472 *Colocalization Analysis*
473
474 For each susceptibility locus of PA (**Table 1**), colocalization analysis was conducted between its
475 most significantly associated phenotype and eQTL effects on gene expression in 48 tissues in
476 GTEx v7[28]. SNPs within +-200kb radius of the lead SNP were used and genes that had at least
477 one significant eQTL (q-value < 0.05) in the region were considered. Analysis was conducted
478 using the R package COLOC [52] and GWAS and eQTL effects were identified as being colocalized
479 if PP4 > 0.8.
480

481 *Heritability and Genetic Correlation Analysis*

482

483 Heritability of activity phenotypes was estimated using Haseman-Elston regression as an
484 intermediate output of fastGWA[19]. Our fastGWA analysis computed sparse GRM at sparsity
485 level 0.05 as recommended by the fastGWA paper (see "Association analysis"). However, this
486 cutoff may miss the subtle relatedness in the sample and affect heritability estimate. As a
487 sensitivity analysis, we re-estimated the heritability using a lower sparsity threshold at 0.02 to
488 capture more subtle relatedness. The genetic correlation between 18 PA traits and 238 complex
489 traits and diseases was estimated using LD score regression[24] implemented in LD Hub[53]. In
490 particular, we focused on four broad groups of traits and diseases (A) cholesterol levels (B)
491 anthropometric traits (C) autoimmune disease and (D) miscellaneous traits including
492 psychiatric, neurological, cognitive and personality traits. For each trait and within each
493 category, we applied a false discovery rate correction to the p-values corresponding to the
494 genetic correlation estimated using LD score regression, to account for multiple testing. Any
495 genetic correlation with FDR-adjusted p-value less that 10% were declared as significant.

496

497 **Data Availability**

498 Data supporting the findings of this paper are available upon application to the UK Biobank
499 study (https://www.ukbiobank.ac.uk/). The summary statistics will be made publicly available
500 upon publication.

501

502 **Acknowledgements**

The UK Biobank data was accessed via application ID 17712. Research of Drs. Guanghao Qi, Diptavo Dutta and Nilanjan Chatterjee was supported by an R01 grant from the National Human Genome Research Institute [1 R01 HG010480-01].

Competing Interests

Dr. Ciprian Crainiceanu is consulting with Bayer and Johnson and Johnson on methods development for wearable devices in clinical trials. The details of the contracts are disclosed through the Johns Hopkins University eDisclose system and have no direct or apparent relationship with the current paper. The other authors declare no conflict of interest.

Ethics Statement

Approval was granted by the UK Biobank study under application ID 17712 to use the data in the present work. UK Biobank has ethical approval from the North West Multi-Centre Research Ethics Committee (approval number 16/NW/0274). All participants provided informed consent to participate.

Web Resources

UK Biobank: https://www.ukbiobank.ac.uk/

fastGWA software: https://cnsgenomics.com/software/gcta/#fastGWA

FUSION TWAS software: http://gusevlab.org/projects/fusion/

525 COLOC R package: https://cran.r-project.org/web/packages/coloc/index.html

526 LD score regression software: https://github.com/bulik/ldsc

527 LD Hub: http://ldsc.broadinstitute.org/ldhub/

528 FUMA GWAS: https://fuma.ctglab.nl/

529 PLINK: https://www.cog-genomics.org/plink/2.0/

530 Molecular Signatures Database: http://www.broadinstitute.org/msigdb

531

532 **References**

533

534 1. Kyu, H. H., Bachman, V. F., Alexander, L. T., Mumford, J. E., Afshin, A., Estep, K., Veerman,
535 J. L., Delwiche, K., Iannarone, M. L., Moyer, M. L., Cercy, K., Vos, T., Murray, C. J. &
536 Forouzanfar, M. H. Physical activity and risk of breast cancer, colon cancer, diabetes,
537 ischemic heart disease, and ischemic stroke events: systematic review and dose-response
538 meta-analysis for the Global Burden of Disease Study 2013. *BMJ* **354**, i3857 (2016).

539 2. Rovio, S., Kåreholt, I., Helkala, E. L., Viitanen, M., Winblad, B., Tuomilehto, J., Soininen, H.,
540 Nissinen, A. & Kivipelto, M. Leisure-time physical activity at midlife and the risk of
541 dementia and Alzheimer's disease. *Lancet Neurol* **4**, 705-711 (2005).

542 3. Smirnova, E., Leroux, A., Cao, Q., Tabacu, L., Zipunnikov, V., Crainiceanu, C. & Urbanek, J.
543 The predictive performance of objective measures of physical activity derived from
544 accelerometry data for 5-year all-cause mortality in older adults: NHANES 2003-2006. *J*
545 *Gerontol A Biol Sci Med Sci* (2019).

546 4. Leroux, A., Di, J., Smirnova, E., Mcguffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L.,
547 Zipunnikov, V., Urbanek, J. K. & Crainiceanu, C. Organizing and analyzing the activity data
548 in NHANES. *Stat Biosci* **11**, 262-287 (2019).
549 5. Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A.,
550 George, S. M. & Olson, R. D. The Physical Activity Guidelines for Americans. *JAMA* **320**,
551 2020-2028 (2018).
552 6. Lightfoot, J. T., DE Geus, E. J. C., Booth, F. W., Bray, M. S., DEN Hoed, M., Kaprio, J., Kelly,
553 S. A., Pomp, D., Saul, M. C., Thomis, M. A., Garland, T. & Bouchard, C. Biological/Genetic
554 Regulation of Physical Activity Level: Consensus from GenBioPAC. *Med Sci Sports Exerc* **50**,
555 863-873 (2018).
556 7. Moore-Harrison, T. & Lightfoot, J. T. Driven to be inactive? The genetics of physical
557 activity. *Prog Mol Biol Transl Sci* **94**, 271-290 (2010).
558 8. De Moor, M. H., Liu, Y. J., Boomsma, D. I., Li, J., Hamilton, J. J., Hottenga, J. J., Levy, S., Liu,
559 X. G., Pei, Y. F., Posthuma, D., Recker, R. R., Sullivan, P. F., Wang, L., Willemsen, G., Yan, H.,
560 De Geus, E. J. & Deng, H. W. Genome-wide association study of exercise behavior in Dutch
561 and American adults. *Med Sci Sports Exerc* **41**, 1887-1895 (2009).
562 9. Hara, M., Hachiya, T., Sutoh, Y., Matsuo, K., Nishida, Y., Shimanoe, C. et al. Genomewide
563 Association Study of Leisure-Time Exercise Behavior in Japanese Adults. *Med Sci Sports*
564 *Exerc* **50**, 2433-2441 (2018).
565 10. Kim, J., Min, H., Oh, S., Kim, Y., Lee, A. H. & Park, T. Joint identification of genetic variants
566 for physical activity in Korean population. *Int J Mol Sci* **15**, 12407-12421 (2014).

567 11. Klimentidis, Y. C., Raichlen, D. A., Bea, J., Garcia, D. O., Wineinger, N. E., Mandarino, L. J.,
568 Alexander, G. E., Chen, Z. & Going, S. B. Genome-wide association study of habitual
569 physical activity in over 377,000 UK Biobank participants identifies multiple variants
570 including CADM2 and APOE. *Int J Obes (Lond)* **42**, 1161-1176 (2018).
571 12. Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van
572 Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T.,
573 Brage, S. & Wareham, N. J. Large Scale Population Assessment of Physical Activity Using
574 Wrist Worn Accelerometers: The UK Biobank Study. *PLoS One* **12**, e0169649 (2017).
575 13. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic,
576 D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G.,
577 Leslie, S., Allen, N., Donnelly, P. & Marchini, J. The UK Biobank resource with deep
578 phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
579 14. Doherty, A., Smith-Byrne, K., Ferreira, T., Holmes, M. V., Holmes, C., Pulit, S. L. & Lindgren,
580 C. M. GWAS identifies 14 loci for device-measured physical activity and sleep duration.
581 *Nat. Commun.* **9**, 5257 (2018).
582 15. Schrack, J. A., Kuo, P. L., Wanigatunga, A. A., Di, J., Simonsick, E. M., Spira, A. P., Ferrucci,
583 L. & Zipunnikov, V. Active-to-Sedentary Behavior Transitions, Fatigability, and Physical
584 Functioning in Older Adults. *J Gerontol A Biol Sci Med Sci* **74**, 560-567 (2019).
585 16. Rock, P., Goodwin, G., Harmer, C. & Wulff, K. Daily rest-activity patterns in the bipolar
586 phenotype: A controlled actigraphy study. *Chronobiol Int* **31**, 290-296 (2014).

587 17. McGregor, D. E., Palarea-Albaladejo, J., Dall, P. M., Stamatakis, E. & Chastin, S. F. M.
588 Differences in physical activity time-use composition associated with cardiometabolic
589 risks. *Prev Med Rep* **13**, 23-29 (2019).
590 18. Young, D. R. & Haskell, W. L. Accumulation of Moderate-to-Vigorous Physical Activity and
591 All-Cause Mortality. *J Am Heart Assoc* **7**, (2018).
592 19. Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M. & Yang, J. A resource-
593 efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**,
594 1749-1755 (2019).
595 20. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. et al. Integrative
596 approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245-
597 252 (2016).
598 21. Finucane, H. K., Reshef, Y. A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A. et al.
599 Heritability enrichment of specifically expressed genes identifies disease-relevant tissues
600 and cell types. *Nat. Genet.* **50**, 621-629 (2018).
601 22. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R. et al.
602 Partitioning heritability by functional annotation using genome-wide association summary
603 statistics. *Nat. Genet.* **47**, 1228-1235 (2015).
604 23. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
605 annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
606 24. Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., ReproGen, C.,
607 Psychiatric, G. C., Genetic, C. F. A. N. O. T. W. T. C. C. C., Duncan, L., Perry, J. R., Patterson,

608 N., Robinson, E. B., Daly, M. J., Price, A. L. & Neale, B. M. An atlas of genetic correlations
609 across human diseases and traits. *Nat. Genet.* **47**, 1236-1241 (2015).
610 25. Rock, P., Goodwin, G., Harmer, C. & Wulff, K. Daily rest-activity patterns in the bipolar
611 phenotype: A controlled actigraphy study. *Chronobiol Int* **31**, 290-296 (2014).
612 26. Jones, S. E., Lane, J. M., Wood, A. R., van Hees, V. T., Tyrrell, J., Beaumont, R. N. et al.
613 Genome-wide association analyses of chronotype in 697,828 individuals provides insights
614 into circadian rhythms. *Nat. Commun.* **10**, 343 (2019).
615 27. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R.
616 J., Eyler, A. E., Denny, J. C., GTEx, C., Nicolae, D. L., Cox, N. J. & Im, H. K. A gene-based
617 association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**,
618 1091-1098 (2015).
619 28. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**,
620 204-213 (2017).
621 29. Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia, Working
622 Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M. J., Price, A. L. &
623 Neale, B. M. LD Score regression distinguishes confounding from polygenicity in genome-
624 wide association studies. *Nat. Genet.* **47**, 291-295 (2015).
625 30. Dashti, H. S., Jones, S. E., Wood, A. R., Lane, J. M., van Hees, V. T., Wang, H. et al. Genome-
626 wide association study identifies genetic loci for self-reported habitual sleep duration
627 supported by accelerometer-derived estimates. *Nat. Commun.* **10**, 1100 (2019).

628 31. Fregeau, B., Kim, B. J., Hernández-García, A., Jordan, V. K., Cho, M. T., Schnur, R. E. et al.
629 De Novo Mutations of RERE Cause a Genetic Syndrome with Features that Overlap Those
630 Associated with Proximal 1p36 Deletions. *Am. J. Hum. Genet.* **98**, 963-970 (2016).
631 32. Lin, A. E., Ebert, G., Ow, Y., Preston, S. P., Toe, J. G., Cooney, J. P., Scott, H. W., Sasaki, M.,
632 Saibil, S. D., Dissanayake, D., Kim, R. H., Wakeham, A., You-Ten, A., Shahinian, A., Duncan,
633 G., Silvester, J., Ohashi, P. S., Mak, T. W. & Pellegrini, M. ARIH2 is essential for
634 embryogenesis, and its hematopoietic deficiency causes lethal activation of the immune
635 system. *Nat. Immunol.* **14**, 27-33 (2013).
636 33. Cohn, R. D., Henry, M. D., Michele, D. E., Barresi, R., Saito, F., Moore, S. A., Flanagan, J. D.,
637 Skwarchuk, M. W., Robbins, M. E., Mendell, J. R., Williamson, R. A. & Campbell, K. P.
638 Disruption of DAG1 in differentiated skeletal muscle reveals a role for dystroglycan in
639 muscle regeneration. *Cell* **110**, 639-648 (2002).
640 34. Sharif, K., Watad, A., Bragazzi, N. L., Lichtbroun, M., Amital, H. & Shoenfeld, Y. Physical
641 activity and autoimmune diseases: Get moving and manage the disease. *Autoimmun Rev*
642 **17**, 53-72 (2018).
643 35. Dhalwani, N. N., O'Donovan, G., Zaccardi, F., Hamer, M., Yates, T., Davies, M. & Khunti, K.
644 Long terms trends of multimorbidity and association with physical activity in older English
645 population. *Int J Behav Nutr Phys Act* **13**, 8 (2016).
646 36. Duggal, N. A., Niemiro, G., Harridge, S. D. R., Simpson, R. J. & Lord, J. M. Can physical
647 activity ameliorate immunosenescence and thereby reduce age-related multi-morbidity.
648 *Nat Rev Immunol* **19**, 563-572 (2019).

649 37. Vancampfort, D., Koyanagi, A., Ward, P. B., Rosenbaum, S., Schuch, F. B., Mugisha, J.,
650 Richards, J., Firth, J. & Stubbs, B. Chronic physical conditions, multimorbidity and physical
651 activity across 46 low- and middle-income countries. *Int J Behav Nutr Phys Act* **14**, 6
652 (2017).
653 38. Peters, H. P., De Vries, W. R., Vanberge-Henegouwen, G. P. & Akkermans, L. M. Potential
654 benefits and hazards of physical activity and exercise on the gastrointestinal tract. *Gut* **48**,
655 435-439 (2001).
656 39. Ciloglu, F., Peker, I., Pehlivan, A., Karacabey, K., Ilhan, N., Saygin, O. & Ozmerdivenli, R.
657 Exercise intensity and its effects on thyroid hormones. *Neuro Endocrinol Lett* **26**, 830-834
658 (2005).
659 40. Hawkins, V. N., Foster-Schubert, K., Chubak, J., Sorensen, B., Ulrich, C. M., Stancyzk, F. Z.,
660 Plymate, S., Stanford, J., White, E., Potter, J. D. & McTiernan, A. Effect of exercise on
661 serum sex hormones in men: a 12-month randomized clinical trial. *Med Sci Sports Exerc*
662 **40**, 223-233 (2008).
663 41. Ennour-Idrissi, K., Maunsell, E. & Diorio, C. Effect of physical activity on sex hormones in
664 women: a systematic review and meta-analysis of randomized controlled trials. *Breast*
665 *Cancer Res* **17**, 139 (2015).
666 42. Alessa, H. B., Chomistek, A. K., Hankinson, S. E., Barnett, J. B., Rood, J., Matthews, C. E.,
667 Rimm, E. B., Willett, W. C., Hu, F. B. & Tobias, D. K. Objective Measures of Physical Activity
668 and Cardiometabolic and Endocrine Biomarkers. *Med Sci Sports Exerc* **49**, 1817-1825
669 (2017).

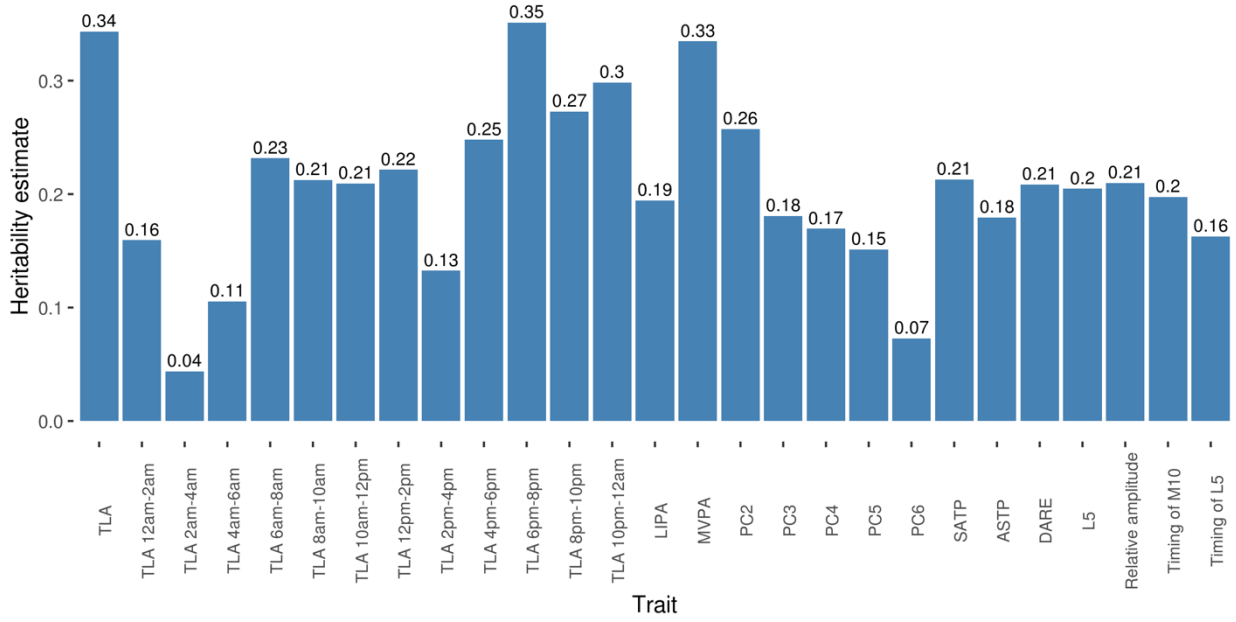
670 43. Hackney, A. C. & Saeidi, A. The thyroid axis, prolactin, and exercise in humans. *Curr Opin*
671 *Endocr Metab Res* **9**, 45-50 (2019).
672 44. Leroux, A., Xu, S., Kundu, P., Muschelli, J., Smirnova, E., Chatterjee, N. & Crainiceanu, C.
673 Quantifying the Predictive Performance of Objectively Measured Physical Activity on
674 Mortality in the UK Biobank. *J Gerontol A Biol Sci Med Sci* (2020).
675 45. Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M.,
676 Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N. & Price, A. L. Efficient
677 Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.*
678 **47**, 284-290 (2015).
679 46. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for
680 biobank-scale datasets. *Nat. Genet.* **50**, 906-908 (2018).
681 47. ENCODE, Project Consortium. An integrated encyclopedia of DNA elements in the human
682 genome. *Nature* **489**, 57-74 (2012).
683 48. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen,
684 A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330
685 (2015).
686 49. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma,
687 D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222
688 (2019).
689 50. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C. et al.
690 The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted
691 arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005-D1012 (2019).

692 51. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. & Mesirov, J.
693 P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).
694 52. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. &
695 Plagnol, V. Bayesian test for colocalisation between pairs of genetic association studies
696 using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
697 53. Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C.,
698 Hemani, G., Tansey, K., Laurin, C., Early Genetics and Lifecourse Epidemiology (EAGLE)
699 Eczema Consortium, Pourcain, B. S., Warrington, N. M., Finucane, H. K., Price, A. L., Bulik-
700 Sullivan, B. K., Anttila, V., Paternoster, L., Gaunt, T. R., Evans, D. M. & Neale, B. M. LD Hub:
701 a centralized database and web interface to perform LD score regression that maximizes
702 the potential of summary level GWAS data for SNP heritability and genetic correlation
703 analysis. *Bioinformatics* **33**, 272-279 (2017).
704
705

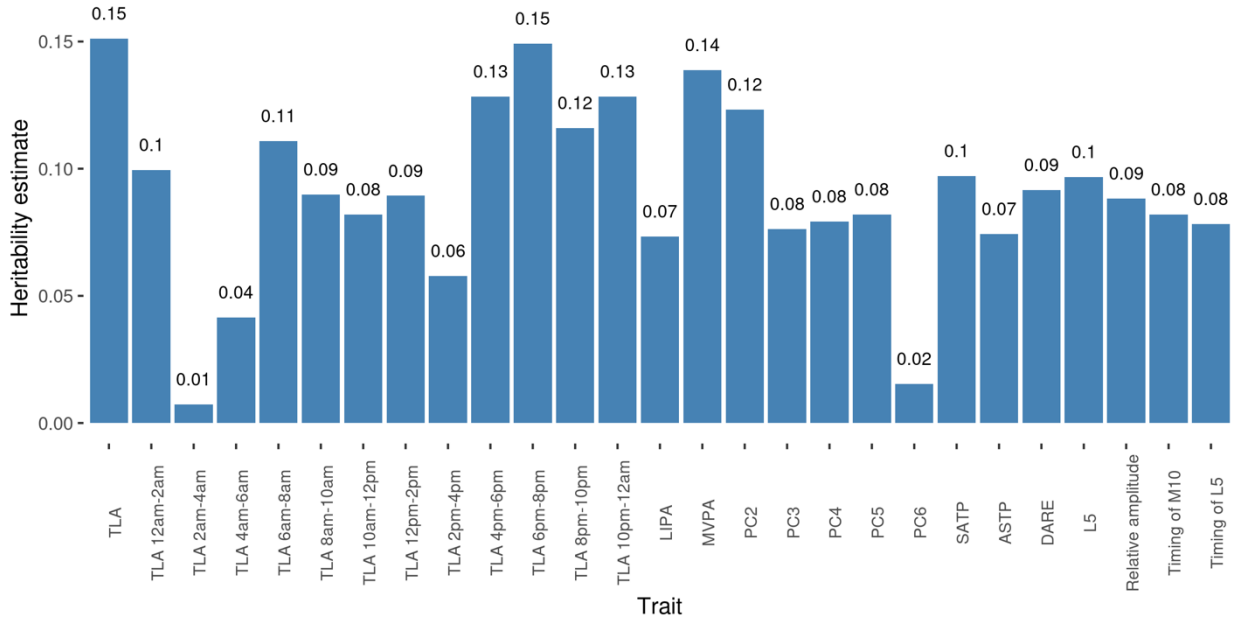
Supplementary Figures

Supplementary Fig. 1 Gene-set enrichment analysis for the significant genes in TWAS analysis for (A) TLA and (B) relative amplitude. We use TLA and relative amplitude as examples because they capture two most important patterns of PA: total volume of activity and sleep quality. Genes previously reported to be associated to different traits and diseases as curated in the GWAS catalog were defined as gene-sets (see Methods). The genes associated to (A) TLA and (B) relative amplitude (p-value $< 2.5 \times 10^{-6}$) in transcriptome-wide analysis were checked for enriched overlap with the gene-sets curated from GWAS catalog for each trait (additionally reported in Molecular Signatures database).

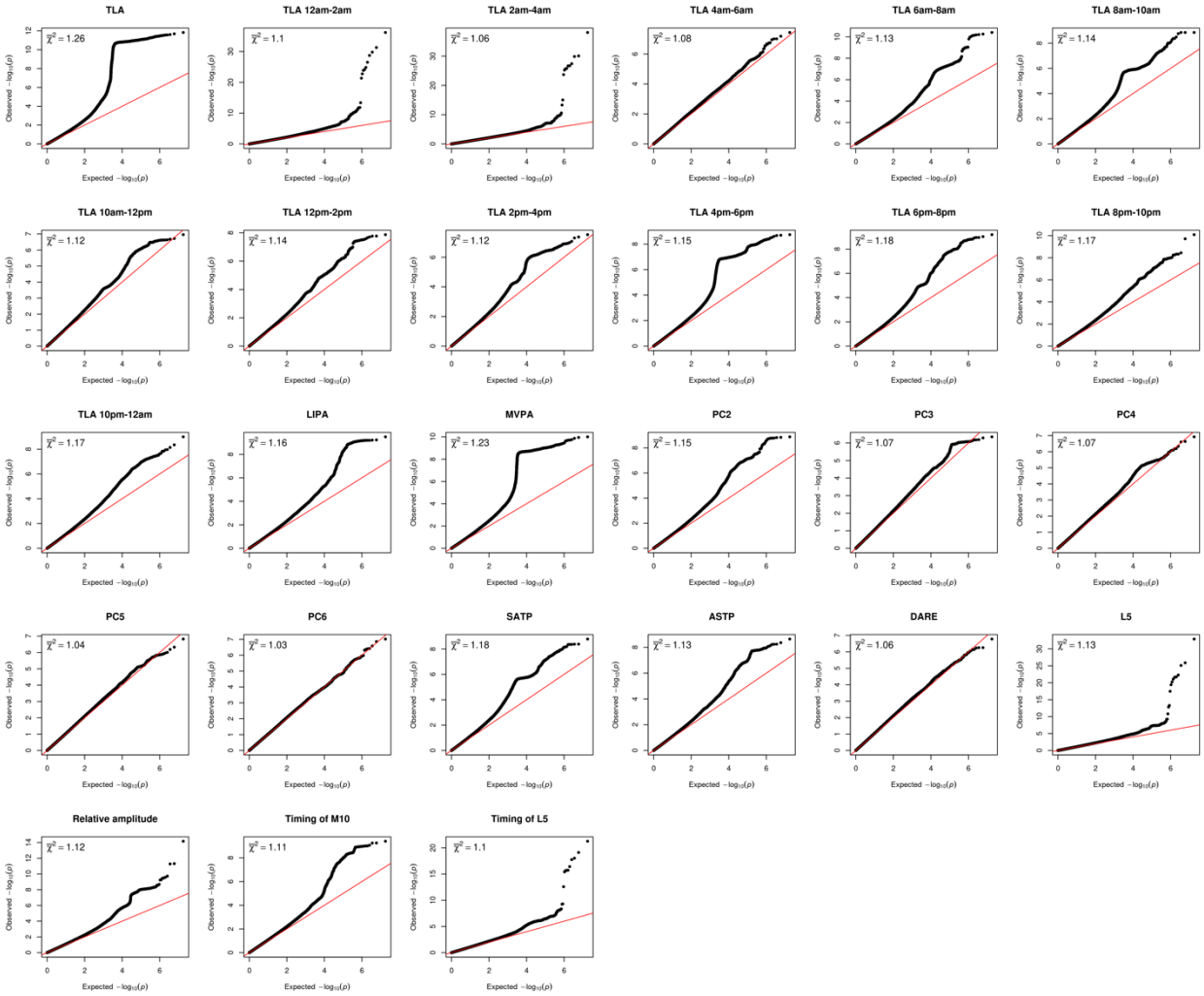
(a) Sparse GRM cutoff = 0.05



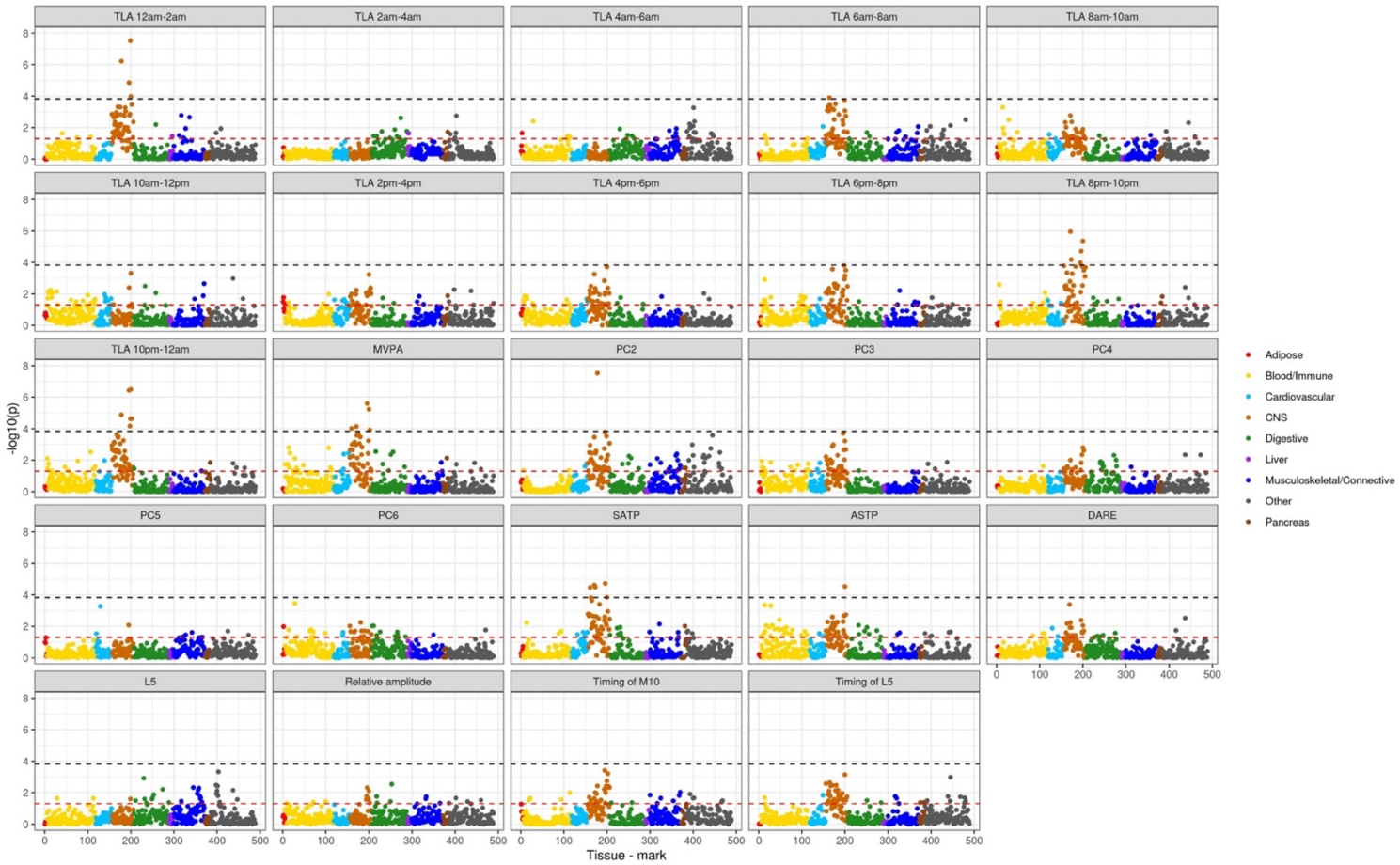
(b) Sparse GRM cutoff = 0.02



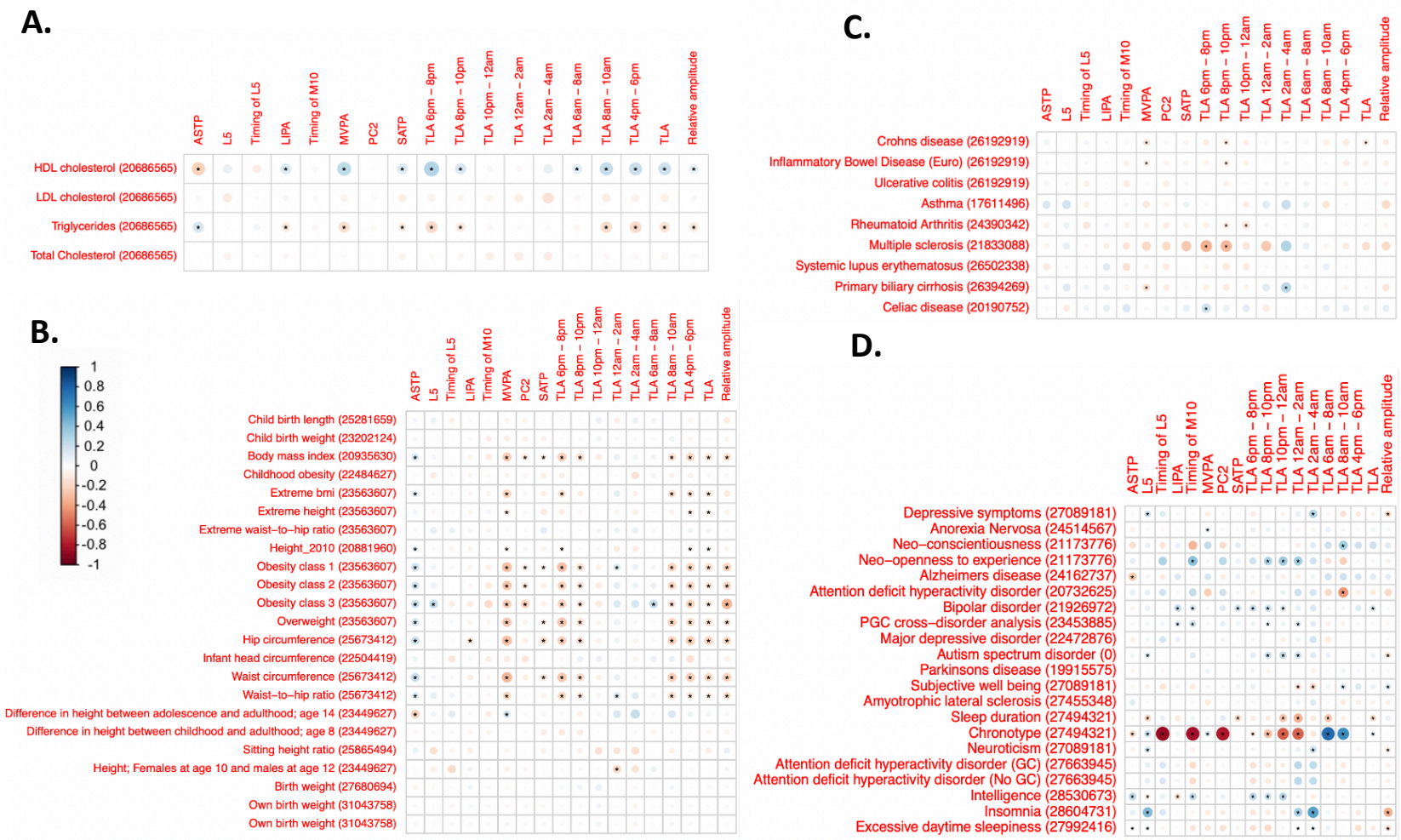
Supplementary Fig. 2 Estimates of heritability of 27 physical activity traits using Haseman-Elston regression and sparse genetic relationship matrix (GRM). a) Cutoff at 0.05: correlations that are < 0.05 in the GRM are reduced to zero, as is in our fastGWA analysis and recommended by the fastGWA paper. b) Cutoff at 0.02: correlations that are < 0.02 in the GRM are reduced to zero.



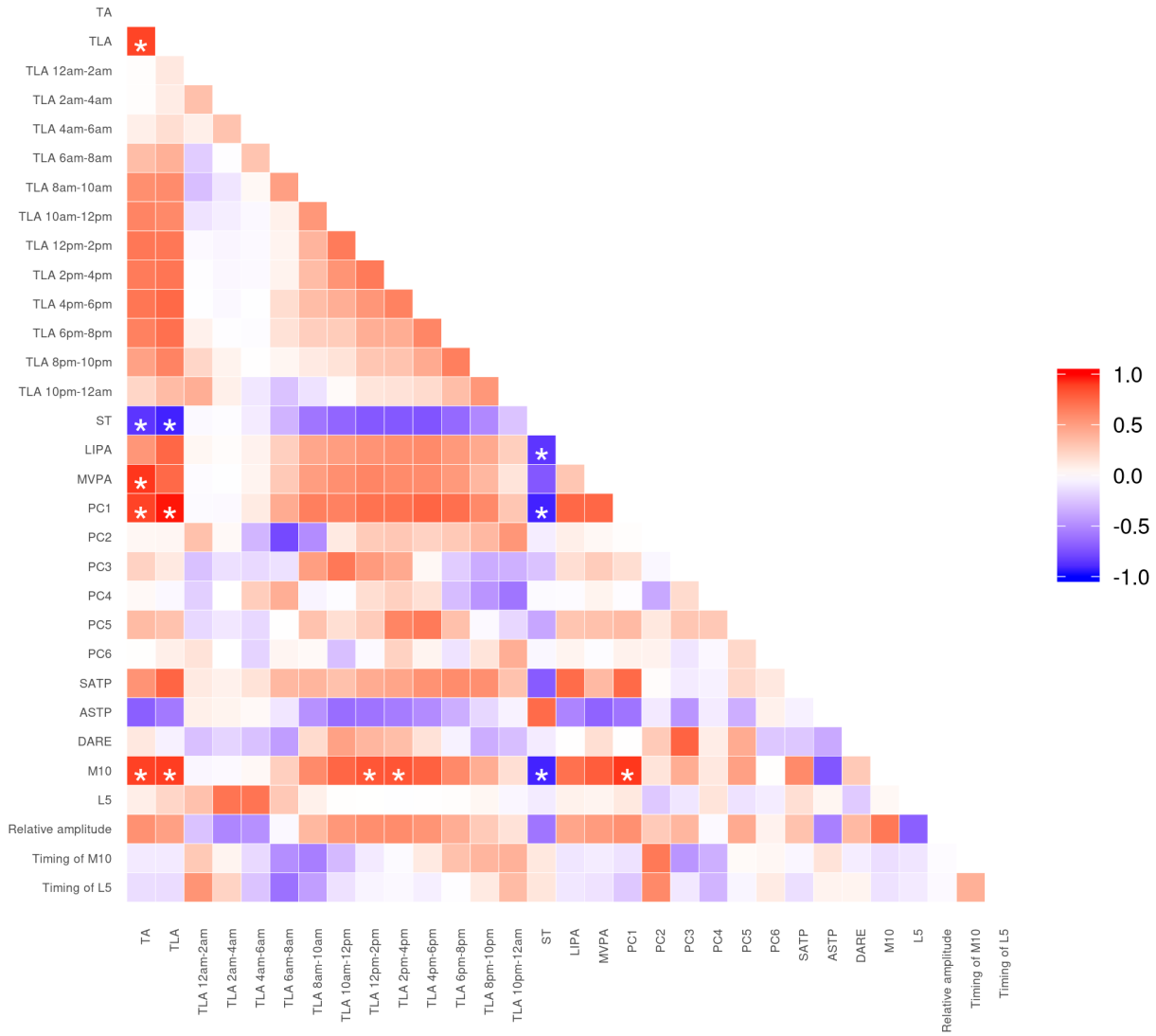
Supplementary Fig. 3 QQ plot for the p-values from fastGWA analysis for 27 activity phenotypes. The average χ^2 statistic is annotated at the top left corner of each panel.



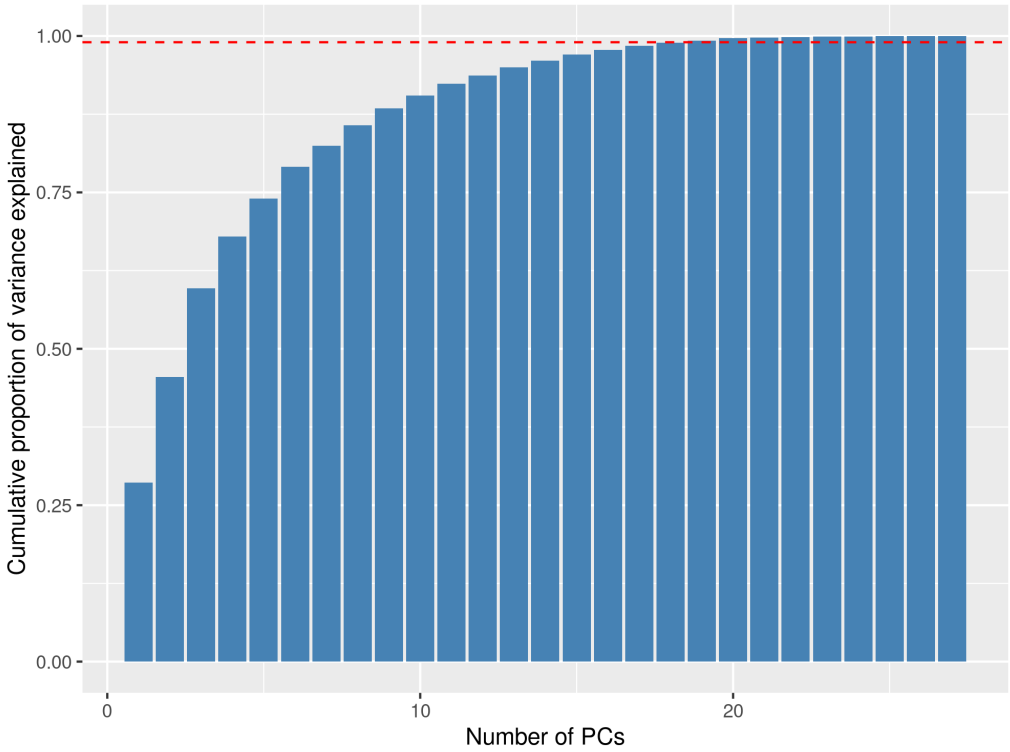
Supplementary Fig. 4 Tissue-specific heritability enrichment for traits not included in Figure 3. The analysis was conducted using tissue/cell type specific stratified LD score regression (Finucane et al, Nat Genet 2018, PMID 29632380) based on 6 chromatin-based annotations in 111 tissues and cell types. A complete list of tissue and cell types is provided in Supplementary Table 7 of the above paper. Black line corresponds to FDR < 0.05 (-log(p-value)=3.83) across all combinations of trait, tissue, and histone mark. Red line corresponds to p = 0.05. The analysis was conducted using tissue/cell type specific stratified LD score regression (Finucane et al, 2018).



Supplementary Fig. 5 Genetic correlation of 18 PA traits with other traits: (A) cholesterol levels (B) anthropometric traits (C) autoimmune diseases and (D) miscellaneous traits including psychiatric, neurological, cognitive and personality traits. The PubMed ID's for the corresponding for the GWAS summary statistics are in parentheses. The size of the circle and the darkness of the color represents the genetic correlation value. The asterisk sign represents if the corresponding genetic correlation was significant at an FDR correction threshold of 10% (corrected within the respective category). For genetic correlation value and p-values across all 237 traits analyzed, see Table S4.



Supplementary Fig. 6 Correlation matrix for 31 physical activity phenotypes. Correlation coefficients > 0.8 or < -0.8 are marked with *.



Supplementary Fig. 7 Number of phenotypic PCs and cumulative proportion of variance explained. At least 19 PCs are needed to explain 99% of the variance (red dashed line).