SNP and Haplotype Regional Heritability Mapping (SNHap-RHM): joint mapping of common and rare variation affecting complex traits

Richard F. Oppong[1,#a*], Thibaud Boutin[2], Archie Campbell[3], Andrew M. McIntosh[4], David Porteous[3], Caroline Hayward[2], Chris S. Haley[2,5], Pau Navarro[2] and Sara Knott[1*]

[1]Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh, United Kingdom.

[2]MRC Human Genetics Unit, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, United Kingdom.

[3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, United Kingdom.

[4]Division of Psychiatry, The University of Edinburgh, Edinburgh, United Kingdom

[5]The Roslin Institute and Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Edinburgh, United Kingdom.

[#a]Current Address: Longitudinal Study Section, Translational Gerontology Branch, National Institute on Aging, National Institutes of Health, Baltimore, Maryland, United States of America

*Corresponding authors: s.knott@ed.ac.uk (SK); oppongrf@nih.gov (RFO)

Short title: SNP and Haplotype-based regional heritability analysis of complex traits

Abstract

We describe a genome-wide analytical approach, SNP and Haplotype Regional Heritability Mapping (SNHap-RHM), that provides regional estimates of the heritability across locally defined regions in the genome. This approach utilises relationship matrices that are based on sharing of SNP and haplotype alleles at local haplotype blocks delimited by recombination boundaries in the genome. We implemented the approach on simulated data and show that the haplotype-based regional GRMs capture variation that is complementary to that captured by SNP-based regional GRMs, and thus justifying the fitting of the two GRMs jointly in a single analysis (SNHap-RHM). SNHap-RHM captures regions in the genome contributing to the phenotypic variation that existing genome-wide analysis methods may fail to capture. We further demonstrate that there are real benefits to be gained from this approach by applying it to real data from about 20,000 individuals from the Generation Scotland: Scottish Family Health Study. We analysed height and major depressive disorder (MDD). We identified seven genomic regions that are genome-wide significant for height, and three regions significant at a suggestive threshold (p-value $< 1 \times 10^{-5}$) for MDD. These significant regions have genes mapped to within 400kb of them. The genes mapped for height have been reported to be associated with height in humans. Similarly, those mapped for MDD have been reported to be associated with major depressive disorder and other psychiatry phenotypes. The results show that SNHap-RHM presents an exciting new opportunity to analyse complex traits by allowing the joint mapping of novel genomic regions tagged by either SNPs or haplotypes, potentially leading to the recovery of some of the “missing” heritability.

Keywords: MDD; height; haplotypes; regional heritability mapping; missing heritability; rare variation; genome-wide analysis

Author Summary

In untangling the genetic contribution to observed phenotype differences, situations can arise where causative variants might be tagged by haplotypes and not in linkage disequilibrium with individual SNPs. This scenario is likely for relatively newly arisen and rarer variants. Here, we propose a regional heritability method, SNHap-RHM, that jointly fits haplotype-based and SNP-based genomic relationship matrices (GRMs) to capture genomic regions harbouring rare variants that the SNP-based GRMs might miss. By analysing ~20,000 Scottish individuals, we show by simulation that the two GRMs are very specific to the type of variant effects they can capture; – the haplotype-based GRMs specifically target haplotype effects which are mostly missed by SNP-based GRMs and vice versa. Applying the method to height and major depressive disorder led to the uncovering of regions in the genome that harbour genes associated with those traits. These results are uniquely important because first they confirm that effects tagged by haplotypes may be missed by conventional SNP-based methods. Secondly, our method, SNHap-RHM, presents an exciting new opportunity to analyse complex traits by allowing the joint mapping of genomic regions tagged by either SNPs or haplotypes, potentially leading to the recovery of some of the “missing” heritability.

65 # Introduction

66 Estimates of the genetic component of complex trait variation using genotyped SNPs
67 led to the conclusion that a proportion of the heritability of complex traits is still unexplained
68 or "missing" (1,2). Full sequence data will contain all the variants that account for all the
69 heritability of complex traits (3). Moreover, some of these true causal variants may be rare
70 (4) and therefore may be in incomplete linkage disequilibrium (LD) with genotyped SNPs (5).
71 Thus, some of the "missing" heritability may be "hidden" in rare variants whose effects are
72 difficult to capture because of lack of statistical power. There is, therefore, some benefit to
73 be gained in terms of improving the heritability estimates and uncovering gene variants
74 involved in the control of traits by fitting genome-wide analytical models that adequately
75 capture the combined effects of rare genetic variants (6,7).

76 In light of this, we proposed a genome-wide analytical approach that draws its
77 theoretical basis from the genome-based restricted maximum likelihood (GREML) approach
78 (1,2,8–10) which utilises both local and genome-wide relationship matrices to provide
79 regional estimates of the heritability across locally defined regions in the genome (11,12). This
80 regional heritability analysis can capture the combined effect of SNPs in a region, and thus
81 small effect variants may be detectable. However, the analysis only captures effects
82 associated with common SNPs present on genotyping chips.

83 Haplotypes may provide a better strategy to capture genomic relationships amongst
84 individuals in the presence of causal rare variants. Although rare variants are not in LD with
85 genotyped variants and thus are difficult to capture in conventional GWAS, these rare
86 variants, may be in LD with some haplotypes and thus can be captured using haplotype
87 methods. Compared with genotyped SNPs, capturing haplotype effects may offer an

88 advantage because haplotypes can be functional units (13). Therefore, haplotype effects may
89 reflect the combined effects of closely linked cis-acting causal variants (14) and using
90 haplotypes could provide real benefit over SNPs in recovering some of the "missing"
91 heritability and identifying novel trait-associated variants. Therefore, we extended the SNP-
92 based regional heritability analysis further by incorporating haplotypes in addition to SNPs in
93 the calculation of the regional GRMs used in the analysis (15). This approach includes two
94 regional GRMs and divides the genome into windows based on local haplotype blocks
95 delimited by recombination boundaries.

96 This paper further explores the properties of both the SNP-based and the haplotype-
97 based regional heritability mapping (SNP-RHM and Hap-RHM respectively). We hypothesise
98 and show by simulation that the Hap-RHM complements existing SNP-RHM analytical
99 approaches by capturing regional effects in the genome that existing SNP-based methods fail
100 to capture. This leads us to propose a mapping strategy that jointly utilises SNP and haplotype
101 GRMs in a single analysis called SNHap-RHM. We then confirm the utility of this approach by
102 applying it to real data obtained from about 20,000 individuals from the Generation Scotland:
103 Scottish Family Health Study (GS: SFHS) (16). We analysed two phenotypes: height and major
104 depressive disorder (MDD). The aim was to uncover novel genetic loci that may affect these
105 traits and improve the estimates of the genetic components of the variation in these traits.

106 Results

107 Overview of methods

108 We have shown previously that regional GREML analysis (Regional Heritability
109 Mapping or RHM) using fixed region sizes in the genome is a suitable mapping method for

110 finding local genetic effects (11). The conventional RHM model fits two genomic relationship
111 matrices (GRMs) in the analyses to map genetic loci that affect trait variation: a local GRM
112 (rGRM) calculated using SNPs located in the region and a genome-wide GRM (gwGRM)
113 calculated from SNPs outside the region. We have since extended this conventional regional
114 heritability analysis to incorporate haplotypes in the calculation of the local GRM and have
115 successfully implemented this in a simulation study (15). This study, like our previous (15),
116 utilises a regional heritability model that breaks the genome into naturally defined regions by
117 delimiting them by recombination hotspots. Two types of regional heritability models are
118 then fitted in turn to the phenotypes. One model (SNP-RHM) uses SNPs to estimate local
119 genetic relationships between study individuals, and the other model (Hap-RHM) estimates
120 local genetic relationships amongst individuals using haplotypes.

121 We first explored the two models in detail using a simulation study in which we
122 simulated 20 replicates of five phenotypes using data from about 20,000 individuals of the
123 GS: SFHS cohort. We then performed a regional heritability analysis that jointly fitted the SNP
124 and the haplotype GRMs in an approach that we termed SNP and Haplotype Regional
125 Heritability Mapping (SNHap-RHM). An overview of SNHap-RHM is shown in Fig 1. We finally
126 applied SNHap-RHM to height and major depressive disorder (MDD) phenotypes of the GS:
127 SFHS.

128 Further details of the models, phenotype simulations and GS: SFHS dataset are
129 presented in the materials and methods section of the manuscript.

130 Simulation study: SNP-RHM, Hap-RHM and SNHap-RHM

131 We performed a regional heritability analysis that fits two GRMs (one for the region
132 and one for the rest of the genome) per region across multiple genomic regions delimited by
133 recombination hotspots (where the estimated recombination frequency exceeds ten
134 centiMorgans per Megabase (10cM/Mb)). This recombination threshold resulted in a total of
135 48,772 regions across the genome. We tested two types of regional heritability models, SNP-
136 RHM and Hap-RHM, on 20 replicates of five simulated phenotypes. In SNP-RHM, the regional
137 matrix is derived from SNP genotypes whereas in Hap-RHM the regional matrix is derived
138 from haplotypes. The phenotypes were simulated to be determined by 20 regional QTL effects
139 and genome-wide polygenic effects. The regional QTL effects of the five phenotypes were
140 simulated using SNPs as causal variants for two of them and haplotypes for the remaining
141 three as described in the methods section.

142 A likelihood ratio test (LRT) was used to test the null hypothesis, H_0: that the genetic
143 variance explained by the region is not significant, against the alternative hypothesis, H_1: that
144 the region accounts for a significant proportion of the phenotypic variance. A large LRT
145 statistic is evidence against the null hypothesis, and therefore means the region explains a
146 significant proportion of the phenotypic variance.

147 The LRTs averaged over the 20 replicates of the five phenotypes are shown in Fig 2.
148 The figure shows plots of average LRT for the QTL regions and ten adjacent regions (5 to each
149 side). The results show that both models detected the simulated regional effects at the
150 genome-wide significance level (LRT = 23.9) (p-value $< 1.02 \times 10^{-6}$, Bonferroni correction
151 for testing 48,772 regions) and can capture true causal loci in traits with different genetic
152 architectures. The LRTs were higher on average for the SNP-based model (SNP-RHM) than the
153 haplotype-based model (Hap-RHM). This could be because for Hap-RHM, the genome-wide

154 GRM which is a SNP-based GRM does not tag any of the background haplotype effects that
155 are outside any one particular region being analysed, and thus the residual variance may be
156 inflated by the other haplotype QTLs which downwardly impact the LRTs.

157 We provide further investigation of the results from the simulation in the supporting
158 information (S1 Text). For both analysis models, we have presented detailed results of the
159 relationships between the LRT statistics, region size, variance estimates and allele frequencies
160 (S3-S10 Figs). We observed that the longer haplotype blocks had many SNPs (and hence many,
161 many haplotypes, up to 14,000 in some blocks), and this impacted the estimation of the
162 simulated regional variance (S8 Fig). We, therefore, performed a hybrid-Hap-RHM analysis
163 that restricted the natural haplotype block sizes to 20 or fewer SNPs per haplotype block. This
164 hybrid-Hap-RHM was to investigate whether the regional variance is well captured by Hap-
165 RHM when shorter haplotypes are used. The hybrid-Hap-RHM underestimated the regional
166 variance for larger regions but did not offer any discernible improvement in the LRT statistics
167 (S9 Fig). The relationship between region size and estimated variance was different between
168 the Hap-RHM and hybrid-Hap-RHM, while we observed a similar relationship between LRTs
169 and the region size.

170 Both SNP-RHM and Hap-RHM fail to capture the simulated regional effects when the
171 simulated phenotype has a genetic architecture that does not match the analysis model, i.e.,
172 SNP or haplotype (Fig 3 and S1 Fig). These figures show the results for the situation where the
173 SNP QTL phenotypes were analysed with the haplotype-based model (Hap-RHM) and the
174 haplotype QTL phenotypes were analysed with the SNP-based model (SNP-RHM). Both
175 models fail to detect the simulated effects in such situations, therefore, showing that the

176 models complement each other since they capture effects due to different types of genetic
177 variants (i.e., tagged by SNPs or haplotypes).

178 To confirm that two models are complementary and independent of each other, we
179 implemented SNHap-RHM that fits the regional SNP and haplotype GRMs jointly, on a
180 replicate of each of the five simulated phenotypes. The significance of regional effects was
181 tested with an LRT with two degrees of freedom. The results are shown in Fig 4 and confirm
182 that the two models are complementary since even when we fitted jointly the two regional
183 matrices (SNP and Haplotype-based), we can still capture the simulated effects.

184 SNHap-RHM analysis of height and MDD in GS: SFHS

185 The heritability estimates for height and MDD in the GS: SFHS dataset, calculated using
186 the whole-genome GRM, were 81.4% (0.92) and 13.8% (1.35) respectively. There were no
187 overlaps between regions identified as significant (tested with an LRT with one degree of
188 freedom) by the haplotype and SNP-based models for either of the two traits (S2 Fig). This
189 reaffirms our hypothesis tested by simulation that the Hap-RHM is complementary to SNP-
190 RHM in mapping associated genomic loci.

191 The regional heritability results for height and MDD are presented as plots of minus-
192 Log10 of the LRT p-values (Figs 5 and 6). The plots for the SNHap-RHM, SNP-RHM and Hap-
193 RHM analyses are shown.

194 The results for height show that nine regions passed the Bonferroni-corrected
195 genome-wide significance threshold in the analysis using SNP-RHM. No region was genome-
196 wide significant for height when analysed with Hap-RHM. Furthermore, seven of the nine
197 associated regions still come up as genome-wide significant when SNPs and haplotypes in

198 those regions are analysed jointly using SNHap-RHM. There are GWAS reported genes that lie
199 in or are within 400kb of these regions (S1 Table).

200 For MDD, no region passed the Bonferroni-corrected genome-wide significance
201 threshold for the analysis done with the SNP-based and haplotype-based regional heritability
202 models (Fig 6). Three regions passed the suggestive significance threshold at p-value $<$
203 1×10^{-5} for Hap-RHM analysis of MDD. A further nine regions were significant at p-value $<$
204 5×10^{-5} for the haplotype-based analysis, and one region for the SNP-based analysis (S2
205 Table). Figure 6 shows that when the two local GRMs are fitted jointly using SNHap-RHM, the
206 genomic regions associated with MDD can still be mapped. The associated regions mapped
207 by the haplotype-based model for MDD contain genes reported by GWAS to be associated
208 with several psychiatry phenotypes (Fig 6 and S2 Table). The most strongly associated region
209 was within 400kb of the *DCC* gene. This gene is part of the NETRIN1 pathway, which has been
210 reported to be associated with major depressive disorder in two GWAS samples (GS: SFHS
211 and Psychiatric Genomics Consortium) (17). Zeng *et al.* (17) used a SNP-RHM guided by
212 pathway analysis (to first uncover pathway association and then localise *DCC* within the
213 pathway) to show the *DCC* association with major depressive disorder. The second most
214 strongly associated region was on chromosome 8, and this region had no gene mapped to it.

215 A linear mixed effects model was used to test for association of the SNPs within the
216 suggestive significant region identified by the haplotype-based model on chromosome 3 for
217 MDD. The model tested for association of SNPs by fitting their allelic dosages individually in a
218 regression model and fitting a GRM to account for relatedness of individuals. The region on
219 chromosome 3 was chosen in this example because there is a psychiatric phenotype
220 associated gene, *MYRIP* (18), mapped to it, unlike the *DCC* region which has the gene outside

221 the region. The results are shown in Table 1. Five SNPs within this region are nominally
222 significant at p-value < 0.05. Four out of these five SNPs confer about 2% increased risk of
223 the disease each. These four SNPs lie within the *MYRIP* gene sequence. The *MYRIP* gene is
224 expressed in the brain (19). A SNP (rs9985399) in this gene is reported to be associated with
225 brain processing speed in the Lothian birth cohort (18). Brain processing speed is an important
226 cognitive function that is compromised in psychiatric illness such as schizophrenia and
227 depression, and old age. Also, a SNP (rs6599077) in the *MYRIP* gene region is associated with
228 sleep duration (20). Sleep durations outside the normal range (both short sleep and long
229 sleep) is significantly associated with increased risk of depression (21–24). The *MYRIP* gene is
230 also reported to have a role in insulin secretion (25) and low insulin levels have been linked
231 to depression (26–28).

232 **Table 1. SNP-based association test of MDD in the *MYRIP* gene region.**

SNP information				Major Depressive Disorder association			
SNP ID	**Chr**	**Pos**	**MAF**	**OR**	**Log (OR)**	**SE (logOR)**	**p**
rs9842160	3	39844703	0.14	0.97	-0.030	0.013	0.02
rs9858242	3	39847606	0.19	1.02	0.025	0.011	0.03
rs1599902	3	39954674	0.41	1.02	0.019	0.009	0.04
rs7618607	3	39947936	0.41	1.02	0.019	0.009	0.04
rs9860916	3	39944942	0.41	1.02	0.019	0.009	0.04

The columns are the SNP ID, chromosome, genome position of SNP, minor allele frequency, odds ratio, log of odds ratio, standard error of log odds ratio and association p-value.

233

234 Comparison with published GWAS SNPs

235 For both traits, the SNPs in the regions that were significant at p-value $< 5 \times 10^{-5}$
236 were compared to SNPs reported in the GWAS catalogue (29) to be significant for the two

traits. The GWAS catalogue was accessed on the 15th of January 2021. The results are presented in Table 2. The SNP-based and haplotype-based models identified 1,380 and 45 SNPs respectively for height, and 78 and 495 SNPs respectively for MDD taking all SNPs within haplotype blocks significant at p-value $< 5 \times 10^{-5}$. Out of the 1,380 SNPs identified for height by the SNP-based model, 57 SNPs spanning 20 haplotype regions were in common with published GWAS results for height.

Table 2. Comparison of SNPs within significant regions identified by both models and published GWAS results for height and MDD.

	Number of SNPS			Number of overlapping SNPS		
Trait	**SNP-RHM**	**Hap-RHM**	**pubGWAS**	**SNP-RHM & Hap-RHM**	**SNP-RHM & pubGWAS**	**Hap-RHM & pubGWAS**
Height	1380	45	4960	0	57	0
MDD	78	495	1815	0	0	0

The columns are the name of trait, number of SNPS in regions identified by SNP-RHM and HAP-RHM with p-value $< \mathbf{5} \times \mathbf{10^{-5}}$ and SNPS in published GWAS (pubGWAS) for the traits, and the number of SNPS overlapping between the three.

Discussion

We have proposed and implemented a genome-wide analytical method that analyses genomic regions using a regional heritability model (11). We have since extended this method to include haplotypes by fitting a regional haplotype-based GRM (Hap-RHM) and redefined genomic regions in our analysis to be delimited by recombination hotspots generated using HapMap Phase II (15,30). In this study, we build on our previous regional heritability methods by exploring the properties of the SNP and haplotype-based regional heritability mapping models by simulation and demonstrate that the two variance components fitted are largely independent of each other (S2 Fig). The novelty in this study is that we show that the two

255 regional matrices fitted in SNP-RHM and Hap-RHM capture two different kinds of effects in
256 terms of genetic architecture, and thus the two variance components can be fitted jointly (by
257 fitting the SNP and haplotype regional matrices together) in a joint marker regional
258 heritability mapping procedure that we call SNHap-RHM.

259 We hypothesised that the Hap-RHM would complement the SNP-RHM. We
260 investigated this hypothesis in a simulation study in which we simulated 20 replicates each of
261 two types of SNP QTL phenotypes and three types of haplotype QTL phenotypes. The results
262 show that the two heritability models can capture the effects of causal variants within
263 genomic loci associated with the phenotype analysed. The results also show that the two
264 models are specific about the type of causal effect they can capture, therefore, providing
265 support for the hypothesis that haplotype-based regional heritability models will complement
266 SNP-based regional heritability models. We provide further support for this hypothesis by
267 fitting the two GRMs jointly and showing (using an LRT with two degrees of freedom) that we
268 can still capture the simulated effects and real effects from real data.

269 We applied SNHap-RHM to height and MDD phenotypes from the Generation
270 Scotland: Scottish Family Health Study. Again, we draw comparisons between the effects
271 captured by the SNP-RHM and the Hap-RHM. The SNP-RHM identified more Bonferroni-
272 corrected genome-wide (GW) significant regions (p-value $< 1.02 \times 10^{-6}$) for height
273 compared to MDD. Fifty-seven of the SNPs identified for height by the SNP-RHM have been
274 reported by other studies to be associated with height. These SNPs spanned 20 genomic
275 regions in the GS: SFHS cohort. Height is a highly polygenic trait with many common genetic
276 variants accounting for most of the additive genetic variation (31). These common genetic
277 variants may be in LD with genotyped SNPs on SNP chips (these chips are disproportionately

278 enriched for common SNPs). Therefore, the SNP-based regional heritability model is better
279 suited for capturing SNP loci in height compared to MDD.

280 MDD is a very heterogeneous phenotype, and thus every MDD case could have a set
281 of genetic and non-genetic risk factors exclusive to them (32). These unique genetic risk
282 factors will mean that a lot of the genetic variants driving the disease will be rare at the
283 population level. Three genomic regions were identified for MDD by the haplotype-based
284 regional heritability model at the suggestive level, p-value $< 1 \times 10^{-5}$. The Hap-RHM works
285 well for MDD because MDD is believed to be driven by rare genetic variants, and the model
286 can capture rare genetic variants. The haplotype model can capture rare variants because of
287 the LD between rare variants (both typed and untyped) and the flanking variants that
288 aggregate to form the haplotypes within the genomic regions. There were no overlaps
289 between regions identified by the Hap-RHM and SNP-RHM for each trait, which again
290 supports the hypothesis that the two models complement each other in mapping associated
291 loci.

292 In both traits, the top significant regions we mapped at p-value $< 5 \times 10^{-5}$ had genes
293 mapped to those regions or within 400kb of those regions. For height, these genes have been
294 reported to be associated with height in humans (33–39). For MDD, these genes have been
295 reported to be associated with major depressive disorder and other psychiatry phenotypes
296 (17,18,40–43). In one of such regions for MDD, five SNPs within the region are individually
297 significantly associated with MDD at the nominal level (p-value < 0.05). Four of these SNPs
298 lie within the gene sequence of *MYRIP*, and they each confer 2% disease risk. A conventional
299 GWAS analysis would have missed these nominally associated SNPs because they will not
300 reach the suggestive significance threshold, let alone genome-wide (GW) significance.

301 However, analysing these SNPs within the region as haplotypes allowed us to detect the
302 combined effect of these SNPs in the region at a suggestive-significance level even with our
303 relatively small sample size compared to recent genome-wide association studies of MDD:
304 322,580 (44) and 480,359 (43).

305 The current study's primary strength is that we show the ability of SNHap-RHM to
306 incorporate SNP and haplotype information jointly to map genomic regions that affect
307 complex traits. This gives SNHap-RHM a uniquely useful role to play in the future of complex
308 traits analysis. The plummeting costs of whole-genome resequencing (45) have shifted
309 research focus in GWA studies towards sequence data analysis (46). Although whole-genome
310 sequence data analysis allows incorporating all the genetic variants that drive the phenotypic
311 variation, there may still be some variants whose individual effects may be too small to be
312 picked up in a conventional GWA analysis. However, regionally analysing sequence
313 information can help overcome this because multiple small-effect variants in a region can add
314 up to a substantial regional effect that can be captured by a regional SNP GRM or tagged by
315 a haplotype GRM. Moreover, by defining haplotype blocks using recombination hotspots,
316 whole-genome information can be summarised naturally without setting an arbitrary number
317 of SNPs, and that facilitates integration and comparison across studies. More so, regional
318 heritability analysis of sequence data would be an efficient way to deal with the burden of
319 multiple testing, which has long been a problem of conventional GWAS.

320 One limitation of the current study is the computation burden of the analyses, which
321 necessitates the pre-correction of the phenotypes with the whole-genome GRM before
322 performing SNHap-RHM. This was a leave-one-chromosome-out step involving 22 separate
323 GREML analyses, each fitting a whole-genome GRM that excluded SNPs from one

324 chromosome (47). For our sample of about 20,000 individuals, the precorrection step reduced
325 the computation time needed to perform GREML analysis at each region by approximately
326 33% (15 minutes) and used about 20% (16 gigabytes) less memory. Although this was done
327 to speed up the analysis, the precorrection step was used as an approximation to account for
328 the background polygenic effects of genetic markers outside each region; this would have
329 been about 48,772 separate GREMLs to account for each region. Also, due to the two degrees
330 of freedom test applied in SNHap-RHM, we observed a slight drop in the significance of the
331 associated regions in both height and MDD when SNHap-RHM was applied to those traits.
332 One option would be to use a less stringent test for SNHap-RHM, effectively testing regions
333 assuming only one degree of freedom so that if only one of the variance components
334 significantly contributed to the phenotypic variance the region would be identified for
335 subsequent formal testing of the individual variance components.

336 Finally, although this study thoroughly evaluates the robustness of SNP and Haplotype
337 RHM using simulation and demonstrates the utility of SNHap-RHM in real phenotype analysis,
338 seeking replication in other cohorts will improve our understanding and, more importantly,
339 demonstrate that the analysis is portable across studies and genotyping platforms.

340 Conclusion

341 We have implemented a regional heritability analysis and undertaken analyses of
342 regions in the genome delimited by recombination boundaries and shown by simulation that
343 haplotype-based GRMs can capture genetic variance that may be missed by conventional
344 SNP-based GRMs. We then applied this method in the analysis of real phenotype data from
345 GS: SFHS. Again, we show that the haplotype-based regional heritability model uncovers
346 associations in regions of the genome that explain genetic variance missed by the SNP-based

347 heritability model. In light of this, we further showed that regional effects can still be captured
348 when the two regional GRMs (SNP and haplotype-based) are fitted jointly: an analytical
349 procedure we termed SNHap-RHM. This SNHap-RHM presents an exciting new opportunity
350 to analyse complex traits by allowing the joint mapping of novel genomic regions tagged by
351 either SNPs or haplotypes, potentially leading to the recovery of some of the "missing"
352 heritability.

353 Materials and Methods

354 Ethics Statement

355 Ethical approval for the GS: SFHS study was obtained from the Tayside Committee on Medical
356 Research Ethics (on behalf of the National Health Service).

357 The general statistical setting of a regional heritability analysis

358 Consider a vector $\boldsymbol{y}$ of phenotype values with length n, the linear mixed-effects model
359 for fitting the effects of genomic region i and background polygenic markers is given as:

360 $$y = X\beta + W_i u_i + Z u_b + e$$

361 where $\boldsymbol{y}$ is a vector of phenotypes, $\boldsymbol{X}$ is a design matrix of fixed effects, and $\boldsymbol{\beta}$ is a vector of
362 fixed effects, $\boldsymbol{W_i}$ is a design matrix relating phenotype measures to genetic markers in region
363 i and $\boldsymbol{u_i}$ is a vector of random genetic effects due to region i assumed to be multivariate
364 normal, $MVN\left(0, \sigma^2_{u_i}\boldsymbol{L_{u_i}}\right)$. $\boldsymbol{L_{u_i}}$ is a relationship matrix calculated using markers (SNPs or
365 haplotypes) in region i: calculated in the subsequent sections as $\boldsymbol{G}$ for the SNP and $\boldsymbol{H}$ for the
366 haplotype-based models. $\boldsymbol{Z}$ is a design matrix for background polygenic effects of markers
367 outside the region i and $\boldsymbol{u_b}$ is a vector of random polygenic effect of genetic markers excluded

368 from region i, assumed to be multivariate normal, $MVN\left(0, \sigma_{u_b}^2 \boldsymbol{B}_{\boldsymbol{u_b}}\right)$. $\boldsymbol{B}_{\boldsymbol{u_b}}$ is a relationship

369 matrix calculated using the markers outside the region i: calculated in the subsequent section

370 in the same way as $\boldsymbol{G}$. And $\boldsymbol{e}$ is a vector of residual effects assumed to be multivariate normal,

371 $MVN(0, \sigma_e^2 \boldsymbol{I})$. $\boldsymbol{I}$ is an identity matrix.

372 Under the model, the vector of phenotypes $\boldsymbol{y}$ is assumed to be normally distributed,

373 $N(\boldsymbol{X\beta}, \boldsymbol{V})$ where the variance is

374 $$V = \sigma_{u_i}^2 \boldsymbol{L}_{\boldsymbol{u_i}} + \sigma_{u_b}^2 \boldsymbol{B}_{\boldsymbol{u_b}} + \sigma_e^2 \boldsymbol{I}$$

375 **SNP-RHM: SNP-based regional heritability model**

376 A SNP-based regional heritability analysis was first reported by Nagamine *et al.* (11).

377 The regional heritability analysis approach we employ here differs from the analysis done by

378 Nagamine *et al.* (11) in the way the regions are defined. That analysis defined local regions by

379 breaking the genome into smaller user-defined windows of p SNPs, which overlapped by q

380 SNPs. Here, however, we define regions based on recombination boundaries in the genome.

381 The regional heritability model fits two genetic relationship matrices (GRMs): one local

382 GRM for the region and a whole-genome GRM for the remaining SNPs in the genome that are

383 outside the region. The GRMs are genomic relatedness matrices calculated as the weighted

384 proportion of the local or genome-wide autosomal SNPs shared identity by state (IBS)

385 between pairs of individuals. The SNP IBS matrices are calculated as follows, following the

386 second scaling factor proposed by VanRaden (48)

387 $$G = \frac{MM'}{m}$$

where m is the total number of r local or b background autosomal SNPs, and $\boldsymbol{M}$ is a matrix of genotype codes for the sampled individuals centred by loci means and normalised by the standard deviation of each locus. $\boldsymbol{M}$ is calculated as follows for individual i at locus j

$$M_{ij} = \frac{(x_{ij} - 2p_j)}{\sqrt{2p_j(1 - p_j)}}$$

where x_{ij} is the genotype code at locus j for individual i and takes the values 0, 1 and 2 for AA, Aa and aa genotypes respectively, p_j is the frequency of allele 'a' at locus j**.** The SNP-based relationship for individuals i and k is therefore calculated as follows

$$G_{ik} = \frac{1}{m} \times \sum_{j=1}^{m} \frac{(x_{ij} - 2p_j)(x_{kj} - 2p_j)}{2p_j(1 - p_j)}$$

Hap-RHM: Haplotype-based regional heritability model

The haplotype-based regional heritability model follows theoretically from the SNP-based analysis and utilises haplotypes instead of SNPs as the genetic markers for the regional analysis. The analysis fits two GRMs, a haplotype-based regional GRM and a SNP-based background genome-wide GRM. The haplotype-based GRM is similar to the SNP-based GRM defined in the previous section. For a locally defined region (haplotype block) containing h haplotype variants, the haplotype-based kinship for individuals i and k is calculated as follows

$$H_{ik} = \frac{1}{h} \times \sum_{j=1}^{h} \frac{(d_{ij} - 2p_j)(d_{kj} - 2p_j)}{2p_j(1 - p_j)}$$

where d_{ij} is the diplotype code (coded as the number of copies of haplotype j) for individual i and takes the values 0, 1 and 2 for the $h_t h_t$, $h_t h_j$, $h_j h_j$ diplotypes respectively where haplotype t is any haplotype other than haplotype j, i.e. $t \neq j$, p_j is the haplotype frequency for haplotype j.

408 ## Phenotype simulations

409 Five phenotypes were simulated using available genotypic information of 20,032
410 individuals from the Generation Scotland: Scottish Family Health Study (16). A total of 593,932
411 genotyped SNPs were used, and missing genotypes were filled in by imputation. A total of
412 555,091 SNPs remained after a QC that removed SNPs of MAF < 0.01 and SNPs that were out
413 of Hardy-Weinberg equilibrium at p-value < 0.000001.

414 The five phenotypes were simulated to have a total variance of 1. This total is
415 composed of 0.6 environmental (residual) variance and genetic variance of 0.4. The genetic
416 variance was partitioned into two components, a polygenic variance of 0.3 and a total QTL
417 variance of 0.1 (20 QTLs, each explaining a variance of 0.005). A common polygenic variance
418 was simulated for all five phenotypes from 20,000 markers randomly selected across the
419 genome. The polygenic variance was simulated to be normally distributed with zero mean
420 and variance of 0.3.

421 For each phenotype, 20 regions (haplotype blocks) were randomly selected, one on
422 each autosome (except chromosomes 6 and 8 because of the unusually high LD in the MHC
423 regions on chromosome 6 and a large inversion on chromosome 8 (49)), to simulate
424 quantitative trait loci (QTL). This gave a total of 20 QTLs for each phenotype. The regions were
425 delimited by natural boundaries: recombination hotspots where the estimated
426 recombination frequency exceeds ten centiMorgans per Megabase (10cM/Mb) with the
427 estimated recombination frequency between boundaries being less than ten centiMorgans
428 per Megabase (10cM/Mb) based on the Genome Reference Consortium Human Build 37 (50).
429 This recombination threshold resulted in a total of 48,772 regions across the genome. The
430 number and type of marker used to simulate the QTL are what defined the five phenotypes.

431 The five phenotypes are, a 1-SNP QTL within the haplotype block, a multiple-SNP (5 SNPs) QTL
432 within the haplotype block, two types of 1-haplotype QTL within the haplotype block (taking
433 either a common or a rare haplotype as causal) and multiple (5) haplotype QTL within the
434 haplotype block. Details of these phenotypes are described below.

435 For the haplotype QTL phenotypes, a haplotype block is treated as a single genetic
436 locus having multiple alleles. Each haplotype variant within a block is considered as an allele
437 of that locus. Each study individual will carry two alleles, or have a diplotype, for each locus
438 or haplotype block. The genotype data used to simulate the phenotypes were phased using
439 SHAPEIT2 (51) to produce the haplotypes for study individuals. The multiple haplotype QTL
440 phenotypes were simulated by randomly sampling two rare haplotypes and three common
441 haplotypes within each haplotype block to give five haplotypes per block. The two types of 1-
442 haplotype QTL phenotypes were simulated by randomly sampling a rare haplotype per
443 haplotype block for one type and for the other type a common haplotype was randomly
444 sampled within each haplotype block. S10 Fig gives an indication of the frequencies for the
445 rare (0.00002 to 0.036) and common haplotype (0.008 to 0.906) randomly sampled to
446 simulate the phenotypes. There is a slight overlap between the frequencies for rare and
447 common haplotypes because the regions had already been randomly selected before
448 proceeding to randomly select rare and common haplotypes in those regions. Which means
449 what is rare in one region may be common in another.

450 The individual marker contribution to the polygenic effect and the QTL effects were
451 calculated as follows

452 $$\sigma_j^2 = 2p_j(1 - p_j)g_j^2$$

453 $$g_j = \sqrt{\frac{\sigma_j^2}{2p_j(1-p_j)}}$$

454 where σ_j^2 is the contribution of a marker to the QTL or polygenic variance, g_j is the effect of
455 a SNP j or haplotype j randomly sampled to have polygenic or QTL effect, p_j is the frequency
456 of haplotype j or the effect allele of the SNP j. For the single marker QTL phenotypes, each
457 QTL explained a variance of 0.005. For the multiple marker QTL phenotypes, each causal
458 variant explained the same variance, with the effects scaled to account for LD in the region
459 so each QTL locus explained a variance of 0.005. For the multiple haplotype QTL effects, the
460 haplotype effects were scaled relative to the inverse of their frequency to give a total variance
461 explained by the region of 0.005.

462 Common environmental effects were randomly sampled for the five phenotypes from
463 a normal distribution $N(0, \sigma_e^2)$ where σ_e^2 is 0.6. This, together with a genetic variance of 0.4,
464 gave a total variance of 1 for each phenotype. The final simulated phenotype for an
465 individual i was then calculated as follows

466 $$y(single\ markers\ per\ QTL\ region)_i = \sum_{j=1}^{20000} x_{ij}g_j + \sum_{j=1}^{20} x_{ij}g_j + e_i,$$

467 $$y(multiple\ markers\ per\ QTL\ region)_i = \sum_{j=1}^{20000} x_{ij}g_j + \sum_{l=1}^{20}\sum_{j=1}^{5} x_{ij}g_j + e_i,$$

468 where x_{ij} is the number of copies of the effect allele of SNP j for individual i (for haplotypes,
469 this is defined as d_{ij}; the number of copies of haplotype j for individual i) and g_j is the effect
470 of haplotype j or SNP j. Twenty replicates were analysed for each of the five phenotypes with
471 a different set of QTL markers sampled for each replicate.

472 ## Analysis of simulated data

473 The five simulated phenotypes were analysed using the two models, the SNP-based
474 regional heritability model (SNP-RHM for the SNP QTL phenotypes) and the haplotype-based
475 regional heritability model (Hap-RHM for the haplotype QTL phenotypes). To test the
476 analytical models' specificity, we applied Hap-RHM to SNP QTL phenotypes and SNP-RHM to
477 the haplotype QTL phenotypes. We also performed a Hap-RHM analysis in which the units of
478 analysis in the haplotype blocks were restricted to regions of 20 or fewer SNPs per haplotype
479 block. This was because we observed that longer haplotype blocks had many SNPs (and hence
480 many, many haplotypes, up to 14,000 in some blocks), and this impacted the estimation of
481 the simulated regional effect. The hybrid Hap-RHM, therefore, investigates whether the
482 regional effect is well captured by the haplotype-based model when shorter haplotypes are
483 used.

484 We estimated the regional genetic variance and polygenic variance using restricted
485 maximum likelihood (REML). For each simulated phenotype, we analysed 220 regions in total
486 to map the 20 simulated QTLs. This involved analysing the region containing the QTL and ten
487 adjacent regions (five in either direction). In this way, we limit the analysis to the regions in
488 the genome with simulated effects, thereby reducing computation time considerably. Also,
489 by analysing neighbouring regions, we are able to explore the precision of estimates of the
490 location of regional effects. We assessed the significance of a region using the Likelihood Ratio
491 Test (LRT). The genome-wide significance threshold was calculated to be LRT = 23.9 (p-value
492 $< 1.02 \times 10^{-6}$) using a Bonferroni correction for testing 48,772 regions.

493 Also, we selected one replicate for each simulated phenotype and performed a
494 regional heritability analysis that jointly fitted the SNP and the haplotype GRM in an approach
495 that we termed SNP and Haplotype Regional Heritability Mapping (SNHap-RHM).

496 # GS: SFHS Data

497 ## Genotyping, quality control and phasing of Generation Scotland: Scottish
498 ## Family Health Study dataset

499 The data from the Generation Scotland: Scottish Family Health Study (GS: SFHS)
500 comprised 23,960 participants recruited from Scotland (16,52). The DNA from about 20,032
501 of the participants had been genotyped using the Illumina HumanOmniExpressExome8v1-2_A
502 chip (~700K genome-wide SNP chip) (16). GRCh37 was used throughout.

503 Quality control excluded SNPs and individuals with a call rate less than 98%, SNPs with
504 minor allele frequency (MAF) less than 1% and SNPs that were out of Hardy-Weinberg
505 equilibrium (p-value < 0.000001). A total of 555,091 autosomal SNPs passed quality control
506 for downstream analysis. Phasing of the GS: SFHS data was done using SHAPEIT2 (51). Best
507 guess haplotypes were used. Haplotype blocks were defined using recombination hotspots
508 with a recombination rate of 10cM/Mb inferred from the Reference Consortium Human Build
509 37 (50). Haplotypes variants within blocks were determined using the phased data.

510 ## Phenotype definition

511 MDD status for GS: SFHS participants was assigned following an initial mental health
512 screening questionnaire with the questions: "Have you ever seen anybody for emotional or
513 psychiatric problems?" or "Was there ever a time when you, or someone else, thought you

514 should see someone because of the way you were feeling or acting?" Participants who
515 answered yes to one or both of the screening questions were further interviewed by the
516 Structured Clinical Interview for DSM-IV (SCID) (53). A total of 18,725 participants (2,603 MDD
517 cases and 16,122 controls) were retained for analysis for MDD. A total of 19,944 participants
518 from the GS: SFHS were analysed for height.

519 **SNHap-RHM of MDD and Height**

520 SNHap-RHM fits jointly, the two types of regional GRMs, SNP-based and haplotype-
521 based, in the analysis of phenotypes (Fig 1). We pre-corrected the phenotypes with the whole-
522 genome GRM before performing SNHap-RHM to speed up the GREML analysis of each block.
523 This pre-correction has previously been shown to speed the regional heritability analysis by
524 Shirali *et al.* (15). This is a leave-one-chromosome-out step (47), which involved 22 separate
525 GREML analyses each fitting a whole-genome GRM that excluded SNPs from one
526 chromosome. The residuals from the pre-correction step were then used in the SNHap-RHM
527 analysis. The models adjusted for sex, age, age^2, and the first 20 principal components
528 calculated from the study participants' genomic relationship matrix (calculated using 555,091
529 autosomal SNPs).

530 The significance of a region was tested with a likelihood ratio test (LRT) with two
531 degrees of freedom which compared a model with three variance components fitted (the two
532 regional variances together with the residual variance) against a model with only the residual
533 variance component fitted. The individual regional variance components in all regions were
534 subsequently tested with an LRT with one degree of freedom which compared a model with
535 three variance components fitted against a model with two variance components fitted (one
536 regional variance component dropped from the model).

The p-values obtained from the LRTs were used to generate genome-wide association plots for each phenotype (equivalent to GWAS Manhattan plots). The genome-wide significance threshold was calculated to be LRT = 23.9 (p-value $< 1.02 \times 10^{-6}$) using a Bonferroni correction for testing 48,772 regions. The suggestive significance threshold of a region was set at an LRT = 19.5 (p-value $< 1 \times 10^{-5}$).

Supporting information

S1 Text. Investigating the SNP-RHM and Hap-RHM with simulated phenotypes.

S1 Fig. Plots of average LRT statistics over replicates of QTL loci across the chromosomes for the 20 simulations of each of the three haplotype QTL phenotypes.

S2 Fig. The two analysis models (SNP-RHM and Hap-RHM) are independent of each other in the analysis of height, and Major depressive disorder

S3 Fig. Plots of LRT statistics against QTL region size for the 20 simulations (not averaged) of each of the two SNP QTL phenotypes

S4 Fig. Plots of LRT statistic against QTL region size for the 20 simulations of each of the three haplotype QTL phenotypes.

S5 Fig. Plots of LRT statistic against estimated regional variance for the 20 simulations of the single SNP QTL phenotype.

S6 Fig. Plots of LRT statistic against estimated regional variance for the 20 simulations of each of the three haplotype QTL phenotypes.

S7 Fig. Plots of region size against estimated regional variance for the 20 simulations of the two SNP QTL phenotype.

558 S8 Fig. Plots of region size against estimated regional variance for the 20 simulations of the
559 three haplotype QTL phenotype.

560 S9 Fig. Plots for the 1-rare haplotype QTL phenotype analysed using Hap-RHM (red points)
561 and a hybrid variant of the Hap-RHM (blue points).

562 S10 Fig. Plots of LRT statistic against QTL marker frequencies.

563 S1 Table. Top genomic regions identified by SNP/ haplotype-based model for Height.

564 S2 Table. Top genomic regions identified by SNP/ haplotype-based model for MDD.

565 Acknowledgements

566 We are grateful to all the families who took part, the general practitioners, and the Scottish
567 School of Primary Care for their help in recruiting them, and the whole Generation Scotland
568 team, which includes interviewers, computer and laboratory technicians, clerical workers,
569 research scientists, volunteers, managers, receptionists, healthcare assistants and nurses. We
570 also acknowledge Eilidh Fummey for coming up with the name for the joint mapping method.

571 Author Contributions

572 Conceived and designed the experiments: RFO PN CSH SK. Provided data: TB AC AMM DP CH.
573 Performed the experiments: RFO. Analysed the data: RFO. Wrote the paper: RFO PN CSH SK.

574 References

575 1. Maher B. Personal genomes: The case of the missing heritability. Nature News. 2008 Nov
576 5;456(7218):18–21.

577 2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing
578 heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747–53.

579 3. Wainschtein P, Jain DP, Yengo L, Zheng Z, TOPMed Anthropometry Working Group T-O for
580 PMC, Cupples LA, et al. Recovery of trait heritability from whole genome sequence data.
581 bioRxiv. 2019 Mar 25;588020.

582 4. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum
583 Genet. 2001 Jul;69(1):124–37.

584 5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain
585 a large proportion of the heritability for human height. Nat Genet. 2010 Jul;42(7):565–9.

586 6. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through
587 whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415–25.

588 7. Gonzalez-Recio O, Daetwyler HD, MacLeod IM, Pryce JE, Bowman PJ, Hayes BJ, et al. Rare
589 Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the
590 Missing Heritability of Complex Traits in Cattle. PLoS ONE. 2015 Dec 7;10(12):e0143945.

591 8. Clarke AJ, Cooper DN. GWAS: heritability missing in action? Eur J Hum Genet. 2010
592 Aug;18(8):859–61.

593 9. Speed D, Hemani G, Johnson MR, Balding DJ. Improved Heritability Estimation from Genome-
594 wide SNPs. The American Journal of Human Genetics. 2012 Jul 12;91(6):1011–21.

595 10. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait
596 Analysis. The American Journal of Human Genetics. 2011 Jan 7;88(1):76–82.

597 11. Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Rudan I, et al. Localising Loci
598 underlying Complex Trait Variation Using Regional Genomic Relationship Mapping. PLoS ONE.
599 2012 Oct 15;7(10):e46501.

600 12. Uemoto Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, Wilson J, et al. The power of regional
601 heritability analysis for rare and common variant detection: simulations and application to eye
602 biometrical traits. Frontiers in Genetics. 2013;4:232.

603 13. Vormfelde SV, Brockmöller J. On the value of haplotype-based genotype-phenotype analysis
604 and on data transformation in pharmacogenetics and -genomics. Nat Rev Genet. 2007
605 Dec;8(12).

606 14. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews
607 Genetics. 2006 Oct;7(10):781–91.

608 15. Shirali M, Knott SA, Pong-Wong R, Navarro P, Haley CS. Haplotype Heritability Mapping
609 Method Uncovers Missing Heritability of Complex Traits. Scientific Reports. 2018 Mar
610 21;8(1):4982.

611 16. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM. Cohort profile: Generation
612 Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential
613 for genetic research on health and illness. Int J Epidemiol [Internet]. 2012;42. Available from:
614 http://dx.doi.org/10.1093/ije/dys084

615 17. Zeng Y, Navarro P, Fernandez-Pujals AM, Hall LS, Clarke T-K, Thomson PA, et al. A Combined
616 Pathway and Regional Heritability Analysis Indicates NETRIN1 Pathway Is Associated With
617 Major Depressive Disorder. Biol Psychiatry. 2017 Feb 15;81(4):336–46.

618 18. Luciano M, Hansell NK, Lahti J, Davies G, Medland SE, Räikkönen K, et al. Whole genome
619 association scan for genetic polymorphisms influencing information processing speed. Biol
620 Psychol. 2011 Mar;86(3):193–202.

621 19. Ganat YM, Calder EL, Kriks S, Nelander J, Tu EY, Jia F, et al. Identification of embryonic stem
622 cell-derived midbrain dopaminergic neurons for engraftment. J Clin Invest. 2012
623 Aug;122(8):2928–39.

624 20. Gottlieb DJ, O'Connor GT, Wilk JB. Genome-wide association of sleep and circadian
625 phenotypes. BMC Med Genet. 2007 Sep 19;8(Suppl 1):S9.

626 21. Mohan J, Xiaofan G, Yingxian S. Association between sleep time and depression: a cross-
627 sectional study from countries in rural Northeastern China. J Int Med Res. 2017 Jun;45(3):984–
628 92.

629 22. Roberts RE, Duong HT. The Prospective Association between Sleep Deprivation and Depression
630 among Adolescents. Sleep. 2014 Feb 1;37(2):239–44.

631 23. Watson NF, Harden KP, Buchwald D, Vitiello MV, Pack AI, Strachan E, et al. Sleep Duration and
632 Depressive Symptoms: A Gene-Environment Interaction. Sleep. 2014 Feb 1;37(2):351–8.

633 24. Zhai L, Zhang H, Zhang D. SLEEP DURATION AND DEPRESSION AMONG ADULTS: A META-
634 ANALYSIS OF PROSPECTIVE STUDIES. Depress Anxiety. 2015 Sep;32(9):664–70.

635 25. Waselle L, Coppola T, Fukuda M, Iezzi M, El-Amraoui A, Petit C, et al. Involvement of the Rab27
636 Binding Protein Slac2c/MyRIP in Insulin Exocytosis. Mol Biol Cell. 2003 Oct;14(10):4103–13.

637 26. Greenwood EA, Pasch LA, Shinkai K, Cedars MI, Huddleston HG. Putative role for insulin
638 resistance in depression risk in polycystic ovary syndrome. Fertility and Sterility. 2015 Sep
639 1;104(3):707-714.e1.

640 27. Pearson S, Schmidt M, Patton G, Dwyer T, Blizzard L, Otahal P, et al. Depression and Insulin
641 Resistance. Diabetes Care. 2010 May;33(5):1128–33.

642 28. Webb M, Davies M, Ashra N, Bodicoat D, Brady E, Webb D, et al. The association between
643 depressive symptoms and insulin resistance, inflammation and adiposity in men and women.
644 PLOS ONE. 2017 Nov 30;12(11):e0187448.

645 29. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of
646 published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan
647 4;45(Database issue):D896–901.

648 30. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation
649 human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851–61.

650 31. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation
651 with imputed variants finds negligible missing heritability for human height and body mass
652 index. Nature Genetics. 2015 Aug 31;47(10):1114–20.

653 32. Levinson DF, Mostafavi S, Milaneschi Y, Rivera M, Ripke S, Wray NR, et al. Genetic studies of
654 major depressive disorder: Why are there no GWAS findings, and what can we do about it? Biol
655 Psychiatry. 2014 Oct 1;76(7):510–2.

656 33. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, et
657 al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008
658 May;40(5):609–15.

659 34. Kichaev G, Bhatia G, Loh P-R, Gazal S, Burch K, Freund MK, et al. Leveraging Polygenic
660 Functional Enrichment to Improve GWAS Power. Am J Hum Genet. 2019 Jan 3;104(1):65–75.

661 35. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of
662 variants clustered in genomic loci and biological pathways affect human height. Nature. 2010
663 Oct 14;467(7317):832–8.

664 36. Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, Campbell A, et al. Exploration of haplotype
665 research consortium imputation for genome-wide association studies in 20,032 Generation
666 Scotland participants. Genome Med. 2017 Mar 7;9(1):23.

667 37. Tachmazidou I, Süveges D, Min JL, Ritchie GRS, Steinberg J, Walter K, et al. Whole-Genome
668 Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. Am J
669 Hum Genet. 2017 Jun 1;100(6):865–84.

670 38. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide
671 association analysis identifies 20 loci that influence adult height. Nat Genet. 2008
672 May;40(5):575–83.

673 39. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of
674 common variation in the genomic and biological architecture of adult human height. Nature
675 Genetics. 2014 Nov;46(11):1173–86.

676 40. Arnau-Soler A, Macdonald-Dunlop E, Adams MJ, Clarke T-K, MacIntyre DJ, Milburn K, et al.
677 Genome-wide by environment interaction studies of depressive symptoms and psychosocial
678 stress in UK Biobank and Generation Scotland. Transl Psychiatry. 2019 Feb 4;9(1):14.

679 41. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-
680 analysis of depression identifies 102 independent variants and highlights the importance of the
681 prefrontal brain regions. Nat Neurosci. 2019 Mar;22(3):343–52.

682 42. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million
683 individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet.
684 2019 Feb;51(2):237–44.

685 43. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide
686 association analyses identify 44 risk variants and refine the genetic architecture of major
687 depression. Nature Genetics. 2018 May;50(5):668–81.

688 44. Howard DM, Adams MJ, Shirali M, Clarke T-K, Marioni RE, Davies G, et al. Genome-wide
689 association study of depression phenotypes in UK Biobank identifies variants in excitatory
690 synaptic pathways. Nature Communications. 2018 Apr 16;9(1):1470.

691 45. Caulfield T, Evans J, McGuire A, McCabe C, Bubela T, Cook-Deegan R, et al. Reflections on the
692 Cost of “Low-Cost” Whole Genome Sequencing: Framing the Health Policy Debate. PLOS
693 Biology. 2013 Nov 5;11(11):e1001699.

694 46. Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and
695 precision with whole genome sequencing data in genome-wide association studies of
696 inflammatory biomarkers. Scientific Reports. 2019 Nov 14;9(1):16844.

697 47. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the
698 application of mixed-model association methods. Nat Genet. 2014 Feb;46(2):100–6.

699 48. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008
700 Nov;91(11):4414–23.

701 49. Amador C, Huffman J, Trochet H, Campbell A, Porteous D, Wilson JF, et al. Recent genomic
702 heritage in Scotland. BMC Genomics. 2015;16(1):1–17.

703 50. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of
704 the human genome. Nature. 2004 Oct 21;431(7011):931–45.

705 51. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and
706 population genetic studies. Nature Methods. 2013 Jan;10(1):5–6.

707 52. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ. Generation Scotland: the
708 Scottish Family Health Study; a new resource for researching genes and heritability. BMC Med
709 Genet [Internet]. 2006;7. Available from: http://dx.doi.org/10.1186/1471-2350-7-74

710 53. First MB, Spitzer RL, Gibbon M, Williams JBW. Structured Clinical Interview for DSM-IV-TR Axis I
711 Disorders, Research Version, Non-patient Edition. New York State Psychiatric Institute; 2002.

712

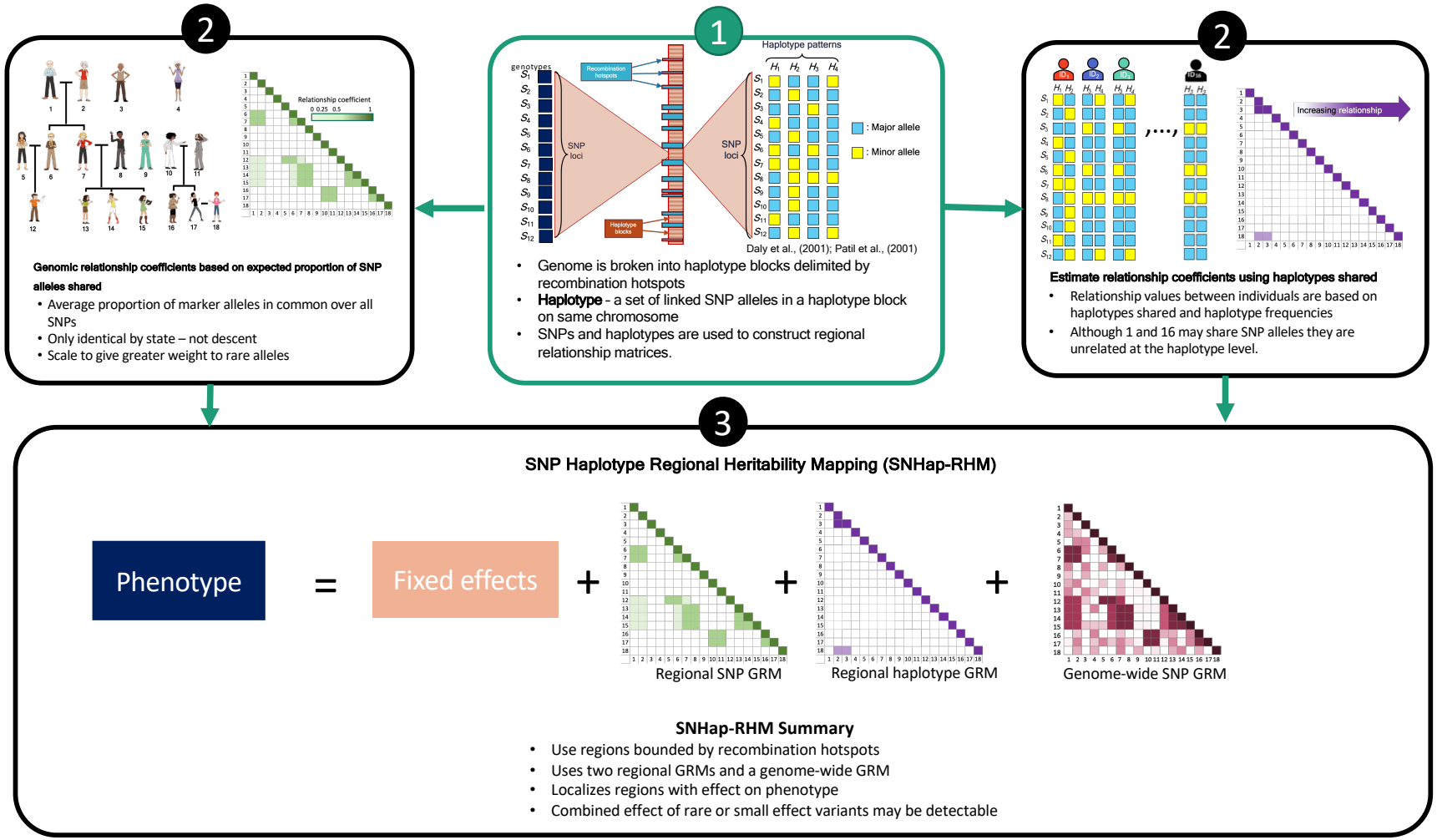

Fig 1. A Schema outlying SNHap-RHM

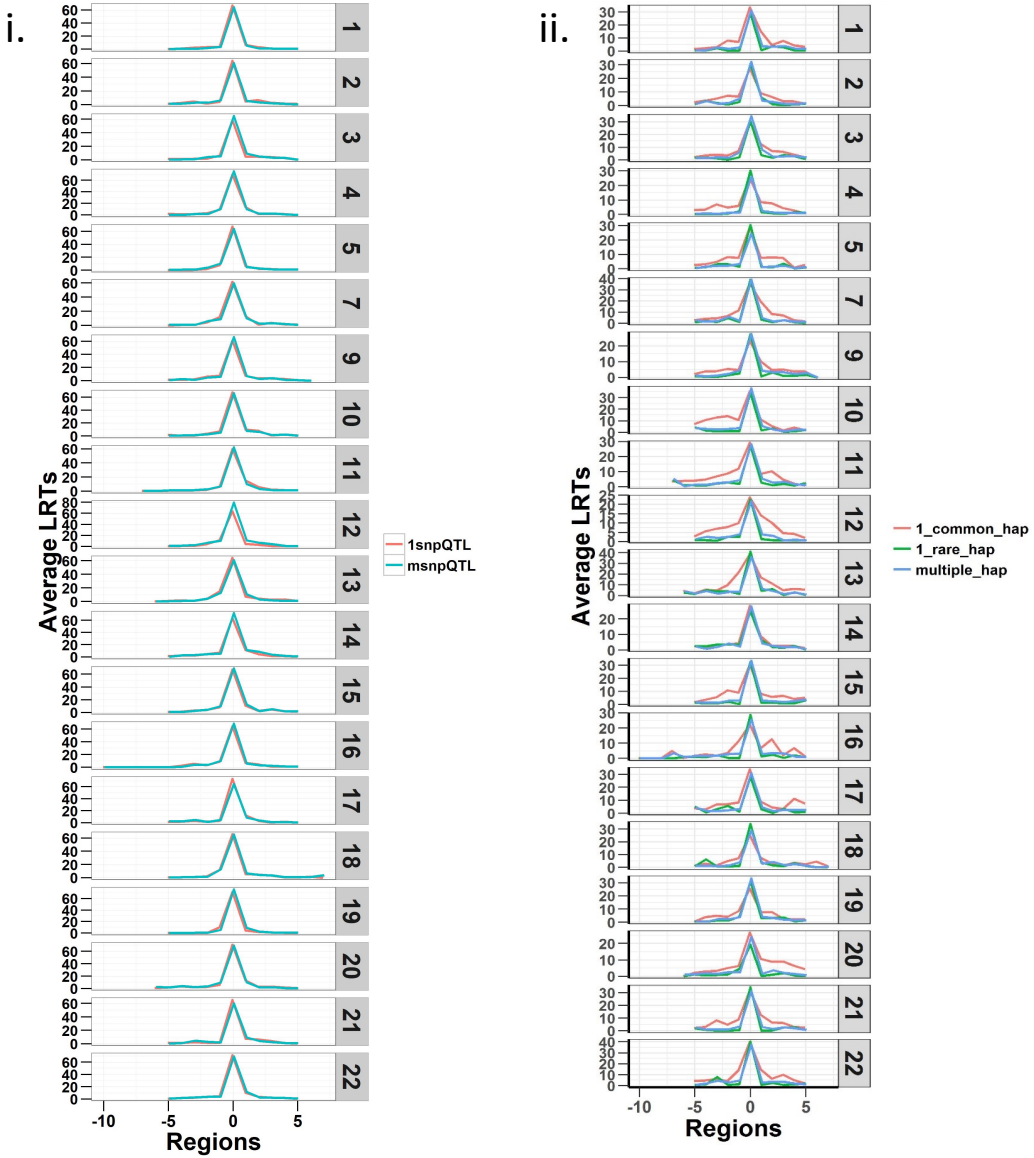

Fig 2. Plots of Likelihood ratio test (LRT) statistics at each QTL locus and 5 regions either side averaged for the 20 simulations of each of the five QTL phenotypes. Plot (i) is SNP QTL phenotypes analysed using the SNP-RHM and plot (ii) is the haplotype QTL phenotypes analysed using the Hap-RHM. Both models can capture the simulated QTL effects for their respective SNP and haplotype phenotypes.

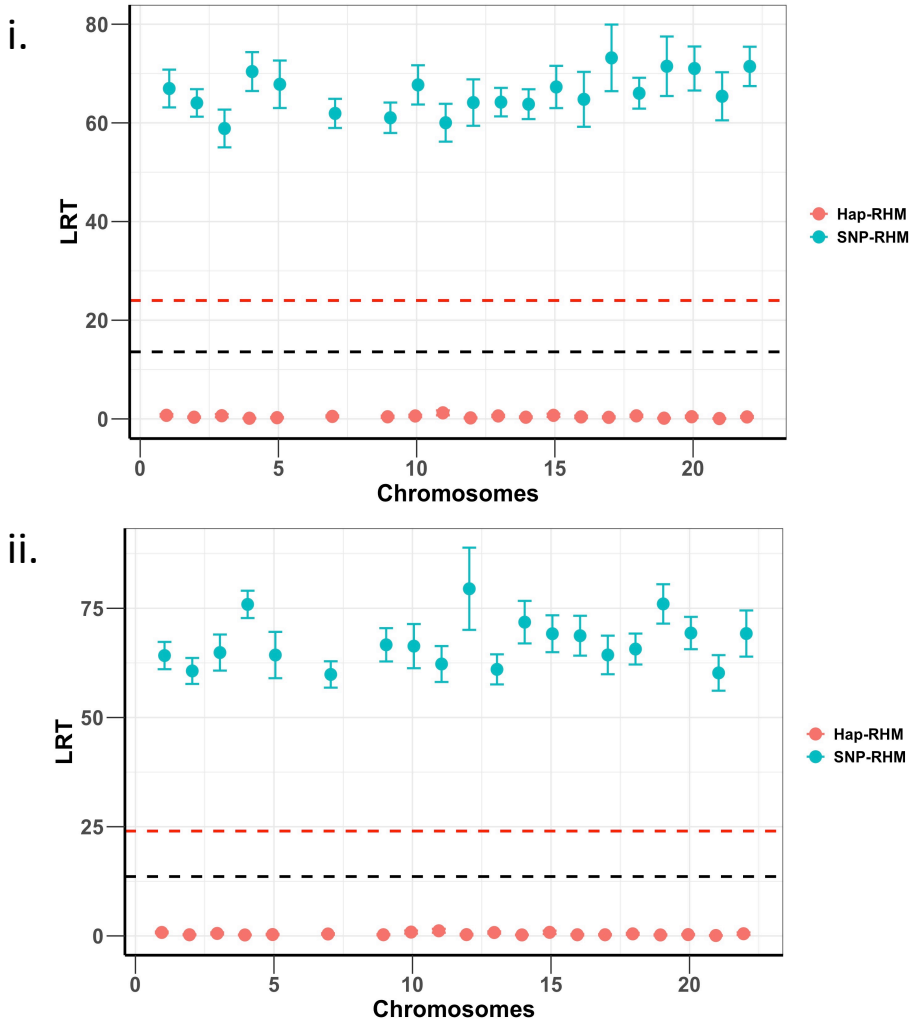

Fig 3. Plots of average LRT statistics over replicates of QTL loci across the chromosomes for the 20 simulations of each of the two SNP QTL phenotypes. The red dashed lines are genome-wide significance threshold (for 48,772 regions) and the black dashed lines are Bonferroni significance threshold (for 220 regions). The upper plot (i) is the 1-SNP QTL phenotype, and the lower plot (ii) is the multiple SNP QTL phenotype. The two phenotypes are analysed using both the SNP based model (SNP-RHM) (blue points) and the Haplotype based model (Hap-RHM) (red points). The Hap-RHM fails to capture the simulated effects for the SNP QTLs.

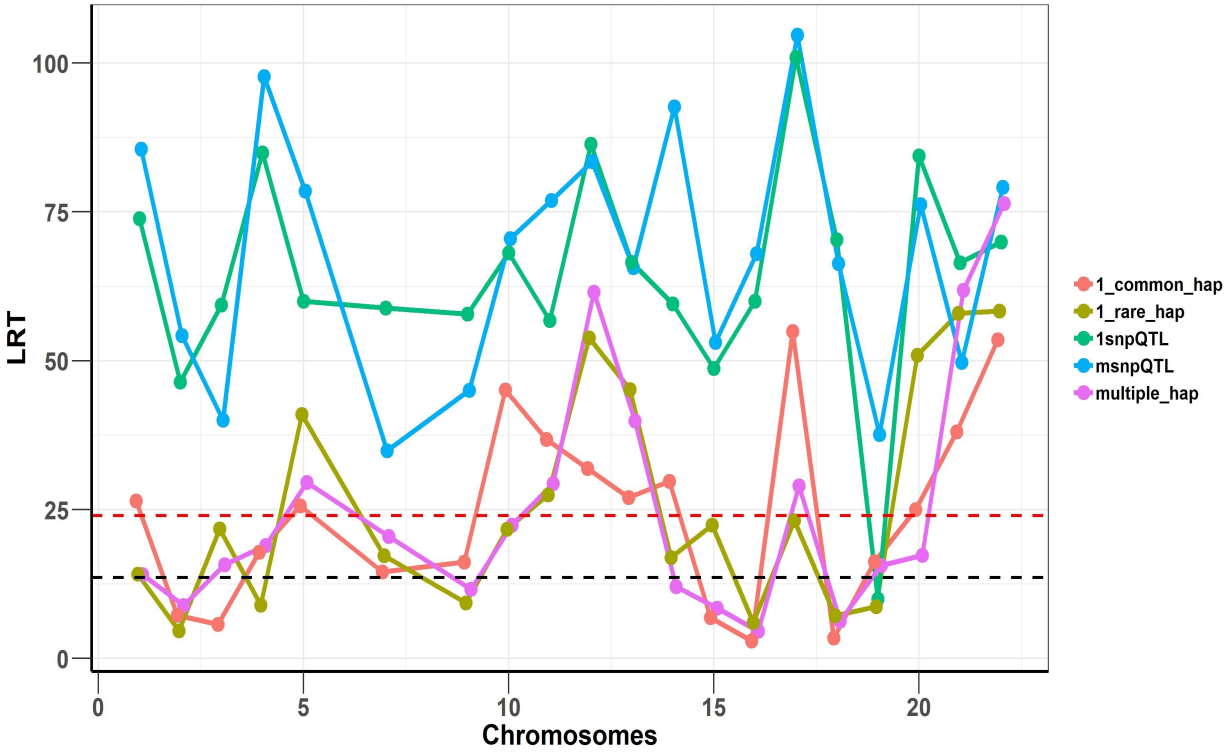

Fig 4. Joint analysis of the SNP and haplotype phenotypes using SNHap-RHM. The plot is an analysis of one replicate of each of the simulated phenotypes. The LRT statistics are plotted over QTL loci across the chromosomes. The red dashed lines are genome-wide significance threshold (for 48,772 regions) and the black dashed lines are Bonferroni significance threshold (for 220 regions).

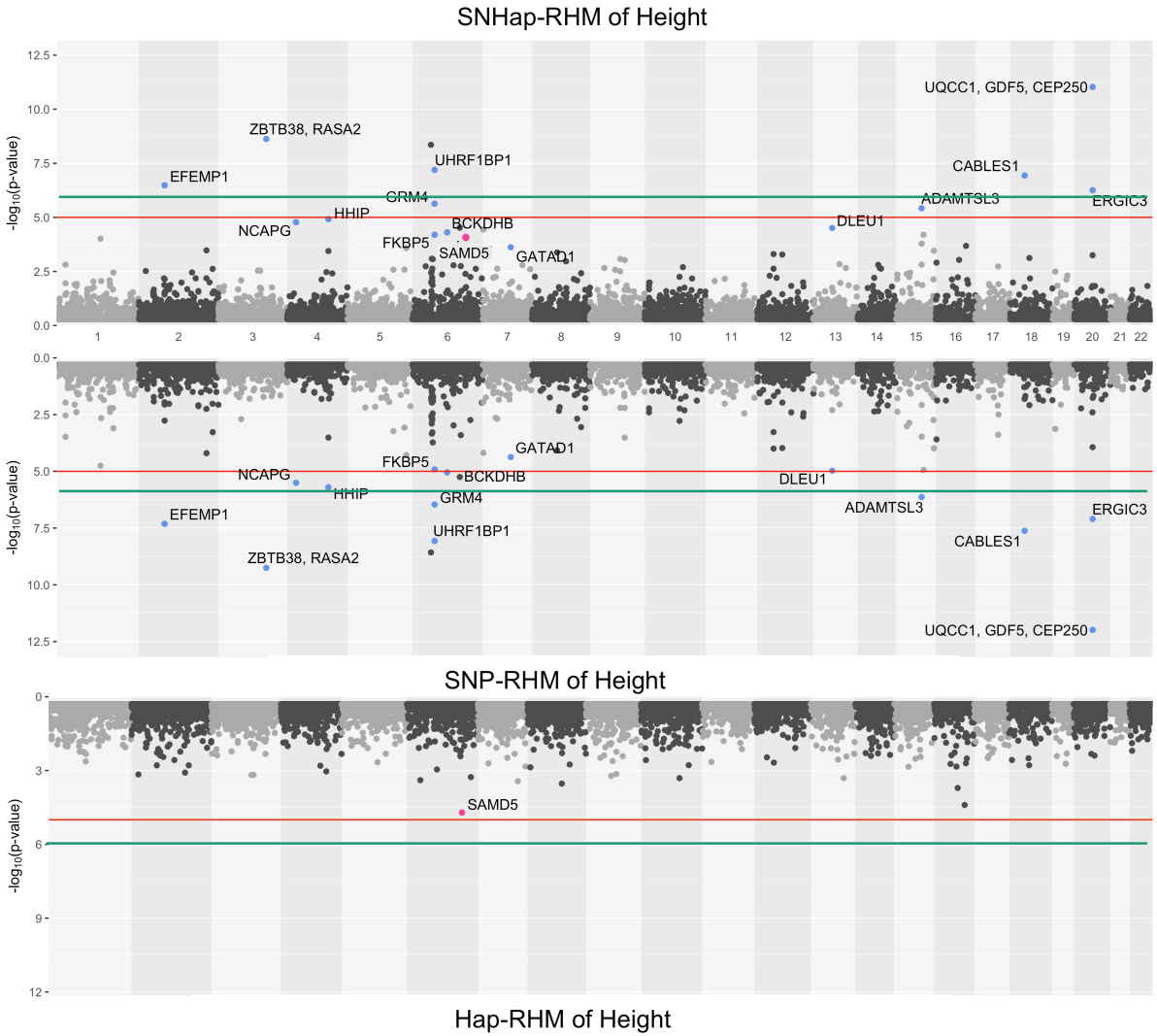

Fig 5. The genome-wide evidence of haplotype block association for height. Analysis done with SNHap-RHM, SNP-RHM and Hap-RHM. The points are plots of -log10 of the p-values of regions tested with the LRT for the regional GREML analyses. The green lines are the Bonferroni-corrected genome-wide significance threshold and the red lines are the suggestive significance threshold calculated to be p-value $< 1 \times 10^{-5}$. The top association hits at p-value $< 5 \times 10^{-5}$ with genes located within the region are highlighted in blue for SNP-RHM and red for the Hap-RHM.

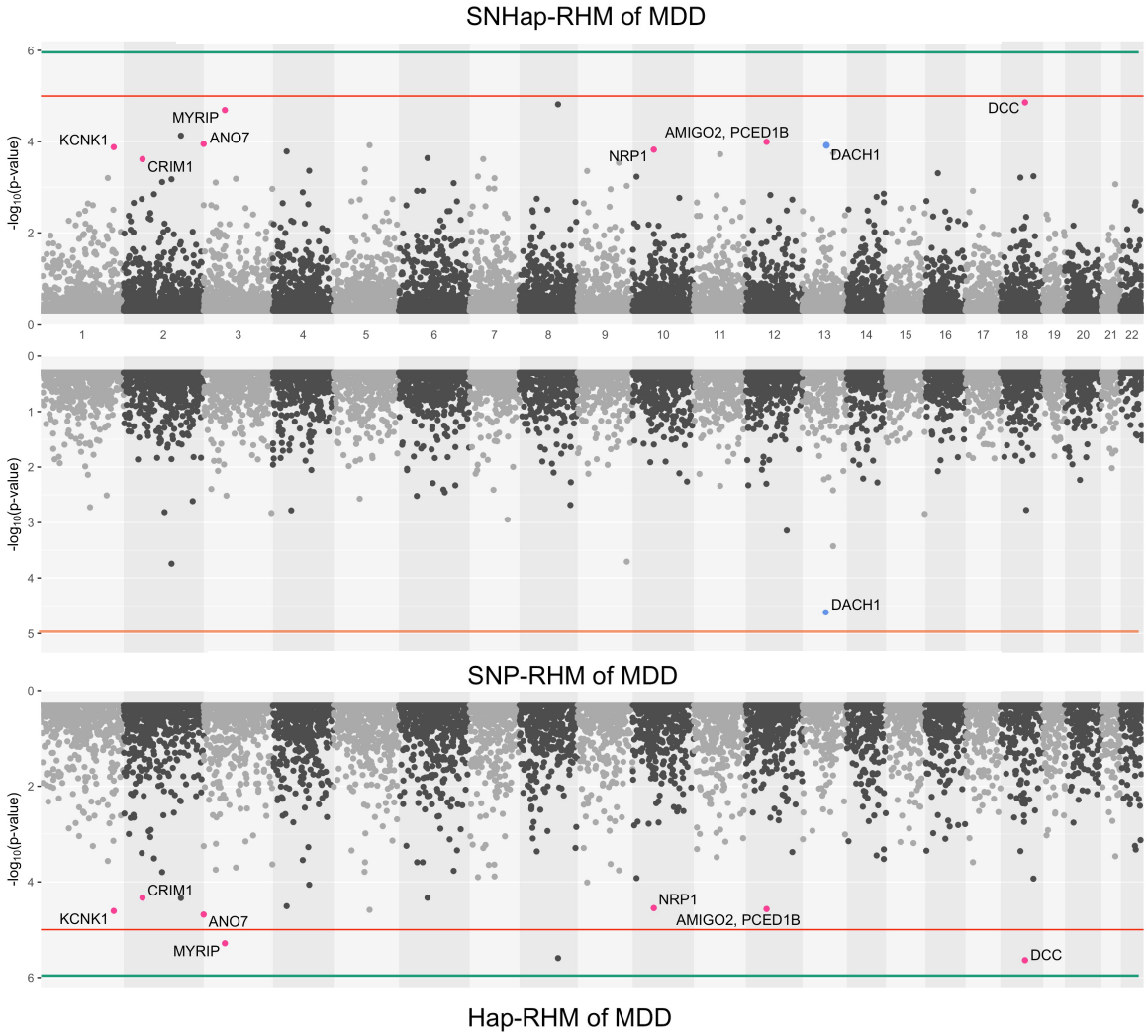

Fig 6. The genome-wide evidence of haplotype block association for Major Depressive Disorder. Analysis done with SNHap-RHM, SNP-RHM and Hap-RHM. The points are plots of -log10 of the p-values of regions tested with the LRT for the regional GREML analyses. The green lines are the Bonferroni-corrected genome-wide significance threshold and the red lines are the suggestive significance threshold calculated to be p-value $< 1 \times 10^{-5}$. The top association hits at p-value $< 5 \times 10^{-5}$ with genes located within the region are highlighted in blue for SNP-RHM and red for the Hap-RHM.