Title: A Genome-Wide Association Analysis of Happiness: Consistent Genetic Effects Across the Lifespan and Across Genetic Ancestries in Multiple Cohorts.

Authors: Joey Ward[1], Laura Lyall[1], Breda Cullen[1], Rona J. Strawbridge[1,2], Xingxing Zhu[1], Ioana Stanciu[1], Alisha Aman[1], Claire L. Niedzwiedz[1], Jana Anderson[1], Mark E. S. Bailey[3], *Donald M. Lyall[1], *Jill Pell[1]

1 Institute of Health & Wellbeing, University of Glasgow, Scotland, UK.

2 Cardiovascular Medicine Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden.

3 School of Life Sciences, University of Glasgow, Scotland, UK.

*These authors contributed equally

Corresponding author

Dr. Joey Ward

Institute of Health and Wellbeing

University of Glasgow

1 Lilybank Gardens

G12 8RZ, Glasgow, UK.

Email: Joey.Ward@Glasgow.ac.uk

Tel: +44 (0) 141 330 8547

Funding

UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. It has also had funding from the Welsh Assembly Government and the British Heart Foundation. The funders had no role in study design, data collection or management, analyses or interpretation of the data, nor preparation, review or approval of the manuscript.

CLN acknowledges funding from the Medical Research Council (MR/R024774/1) and a Lord Kelvin/Adam Smith Fellowship.

Contributions

Concept and design: Joey Ward

Acquisition, analysis, or interpretation of data: Joey Ward, Donald Lyall (imaging), Jill Pell

Drafting of the manuscript: Joey Ward

Critical revision of the manuscript for important intellectual content: all co-authors.

Statistical analysis: Joey Ward

Administrative, technical, or material support: all co-authors.

Conflict of interest statement

The authors have no conflicts of interest to report.

Abstract

We present a genome-wide association study of a general happiness measure in 118,851 participants from the UK Biobank. Using BOLT-LMM, we identify 3 significant loci with a heritability estimate of 0.8%. Linkage disequilibrium score regression was performed on the 'big five' personality traits finding significant associations with lower neuroticism and higher extraversion and conscientiousness. Using a novel approach, we construct LDpred-inf polygenic risk scores in the Adolescent Brain Cognitive Development (ABCD) cohort and the Add Health cohort. We detected nominally significant associations with several well-being measures in ABCD and significant correlations with a happiness measure in Add Health. Additionally, we tested for associations with several brain regions in a white British subsample of UK Biobank finding significant associations with several brain structure and integrity phenotypes.

We demonstrated a genetic basis for general happiness level and brain structure that appears to remain consistent throughout the lifespan and across multiple ancestral backgrounds.

Keywords: Happiness; genetics; GWAS, imaging; bias; UK Biobank, ABCD, Add Health.

Author summary

At the genetic level, there has been little investigation into whether people may have a baseline happiness level which varies from person to person. Here we perform a genetic analysis in the UK Biobank to identify three genetic loci that associate with general happiness level and preform genetic correlations of our results with the 'Big Five' personality traits, identifying significant correlations with neuroticism, conscientiousness and extraversion.

We use the resulting summary statistics to create LDpred-inf polygenic risk scores in UK biobank identifying several brain metrics and regions associate with genetic loading for general happiness level. We also use a novel method to create LDpred-inf polygenic risk scores in two other cohorts, ABCD and Add Health. We found significant correlations with an independent happiness measure in Add Health and nominally significant correlations with several well-being measures in ABCD in both those of European Ancestry and all other ancestries found in these cohorts. We also attempted to replicate our UK Biobank MRi finding in ABCD.

We conclude there is evidence that individuals have a general happiness level that is in part genetic which spans across age and ancestry.

Introduction

Happiness is the core positive emotional state. As an emotional trait that is affected in clinical outcomes (e.g. lack of happiness in depression and exacerbation of happiness in the manic phase of bipolar disorder) it can be considered a positive valence Research Domain Criteria (RDOC) trait[1]. At the genetic level it is usually analysed as part of a wider concept of mental well-being[2]. There is evidence that, generally, an individual has a baseline happiness level which remains relatively stable over time[3] even after major positive or negative life events such as winning the lottery or being the victim of an accident[4].

There has been little investigation into the genetics of general happiness level and most of the research has been performed via twin studies[5], with a recent meta-analysis giving an estimate for heritability of a 'well-being' trait of 36% and a slightly lower estimate of 32% for heritability of 'life satisfaction' [6]. However, to date there has been no genome-wide association study (GWAS) of a general happiness measure to establish features of its genetic

architecture and to identify specific areas of the genome that contribute to this trait. Greater understanding of the genetic basis of subjective happiness could be useful in identifying whether and how targeted interventions could improve happiness, both in the general population and in clinical populations[5]. This in turn will increase our understanding of psychology and neurodevelopment aiding in the future treatment of patients[7].

Here we report the first GWAS of general happiness in individuals of white British ancestry, as self-reported in the UK Biobank[8] cohort, using BOLT-LMM[9] and the use of polygenic risk scores (PRS) in two additional cohorts: the Adolescent Brain Cognitive Development (ABCD) cohort[10] and the National Longitudinal Study of Adolescent to Adult Health (Add Health)[11]. This study had three aims: (1) to identify specific loci in the genome that are associated with self-reported general happiness level; (2) to test whether increased genetic loading of this measure is significantly associated with happiness and well-being measures in independent cohorts that span differing age ranges and different ancestries compared to the discovery GWAS cohort; and (3) to investigate whether genetic predisposition for general happiness level is associated with average differences in brain structure.

Results

Phenotype stability

We established stability of general happiness level in UK Biobank over time in those with a repeat measure (Pearson's r = 0.58, S.E. = 0.01, $p < 2.2 \times 10^{-16}$, n = 4,703).

Weighted Pearson's correlations were calculated to establish the stability of the Add Health happiness phenotype using the function weightedCorr of the R package wCorr, which showed a correlation between waves I and II of 0.34, S.E. = 0.008, $p < 0.0001$ (weighted by wave 2

variable gswgt2) and between waves II and IV of 0.22, S.E. = 0.01, p < 0.0001 (weighted by wave 4 variable gswgt4).

GWAS in UK Biobank

After all exclusions, the final sample size of the GWAS was 118,851, main due to low response rate of the question. The sample had a mean age of 57.5 (S.D. = 8) years and 53.9% were female. A full breakdown of the demographics at each happiness response level can be found in Table S1.

Three loci that achieved genome-wide significance (Table 1, Figure 1, Figures S1-3). They were rs3129791 on chromosome six, rs72743916 on chromosome nine and rs11474549 on chromosome twenty. The heritability estimate h_2^{SNP} was 0.0084 (S.E. = 0.004) and there was minimal inflation of the test statistics as λ_{GC} = 1.001. The LDSR intercept was 0.999 (S.E. = 0.004) showing there was no inflation of the test statistics due to unaccounted population stratification.

Table 1 Genomic loci significantly associated with general happiness level.

SNP	CHR	BP	A1	A2	BETA	SE	P	start of locus (BP)	end of locus (BP)
rs3129791	6	28954293	A	G	-0.023	0.0042	3.70E-08	25038442	32963948
rs72743916	9	98789019	T	C	-0.064	0.011	1.40E-08	97986571	99687691
rs11474549	20	45833129	TCA	T	-0.016	0.0029	2.70E-08	45722939	45930747

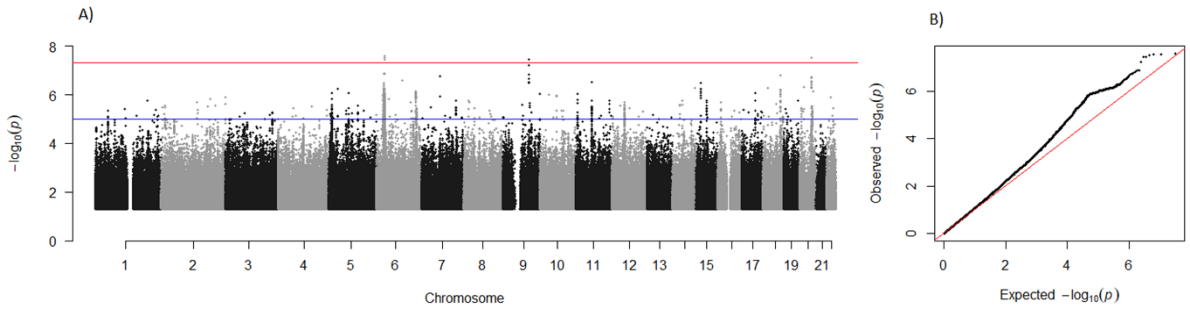

Figure 1 Manhattan plot and QQ plot of general happiness. A) X axis = chromosome from 1 -22 y axis = -log10 p value. Blue line shows suggestive genome wide significance threshold, red line shows genome wide significance threshold. B) x axis = -log10 of the expected p values, y axis = -log10 of the observed p values.

Linkage Disequilibrium Score Regression

Three of the five tested traits were significant after FDR correction (Table 2). The strongest correlation was with neuroticism which showed an inverse relationship with happiness (rg = -0.84, S.E. = 0.34, p.fdr = 0.021). Both extraversion and conscientiousness showed positive correlations with happiness (extraversion: rg = 0.8, S.E. = 0.32, p.fdr = 0.021, conscientiousness: rg = 0.58, S.E. = 0.16, p.fdr = 0.0007).

Table 1 Linkage disequilibrium score regression with general happiness level and the 'Big Five' personality traits.

Trait	rg	se	z	p	P.FDR
conscientiousness	0.589	0.1549	3.80	1.43E-04	7.15E-04
neuroticism	-0.8448	0.33725	-2.51	1.22E-02	0.021
extraversion	0.8025	0.32086	2.50	1.24E-02	0.021
openness	-0.1909	0.10298	-1.85	6.38E-02	0.080
agreeableness	0.8574	0.9374	0.91	3.60E-01	0.360

Polygenic risk scoring

UK Biobank

In those who were excluded from the GWAS due to them having MRI data but who responded to the happiness question and passed genetic quality control (n = 18,795), we detected a significant association of happiness with the LDpred polygenic risk scores (β = 0.057, S.E. = 0.005, p = $4.85*10^{-31}$) with the model explaining 0.018% of the variance.

ABCD

In the white subsample, two of the well-being outcomes tested were significantly correlated with the LDpred happiness PRS (Table 3) but neither survived FDR correction. The strongest associations were with being interested (β = 0.0311, S.E. = 0.016, p = 0.05, p.FDR = 0.13) and being able to concentrate (β = 0.0307, S.E. = 0.015, p = 0.04, p.FDR = 0.12). All of the models

showed that greater genetic loading for happiness was associated with greater agreement with the well-being items.

Table 2 Association of LDpred-inf PRS of general happiness level and well-being measures in ABCD white subsample

outcome	BETA	S.E	P	P.FDR
attentive	8.13E-03	0.017	0.63	0.67
delighted	0.010	0.015	0.51	0.67
calm	0.008	0.014	0.57	0.667
relaxed	0.000	0.014	0.99	0.99
enthusiastic	0.014	0.016	0.39	0.585
interested	0.031	0.016	0.05	0.13
confident	0.027	0.015	0.07	0.15
energetic	0.009	0.016	0.55	0.679
concentrate	0.031	0.015	0.04	0.12

In the whole sample, five of the nine aspects of well-being tested were nominally significant; however, none survived FDR correction (Table 4). The strongest association was with being confident (β = 0.03, S.E. = 0.013, p = 0.018, p.FDR = 0.07). Being enthusiastic and energetic showed the same strength of effects (both: β = 0.029, S.E. = 0.014, p = 0.039, p.FDR = 0.07) followed by being able to concentrate (β = 0.028, S.E. = 0.013, p = 0.028, p.FDR = 0.07) and being calm (β = 0.025, S.E. = 0.012, p = 0.035, p.FDR = 0.07). As with the white subsample analysis, all the models showed that greater genetic loading for happiness associated with greater agreement with the well-being items.

Table 3 Association of LDpred-inf PRS of general happiness level and well-being measures in ABCD multi ancestry sample

outcome	BETA	S.E	P	P.FDR
confident	0.030	0.013	0.018	0.0702
enthusiastic	0.029	0.014	0.039	0.0702
energetic	0.029	0.014	0.039	0.0702
concentrate	0.028	0.013	0.028	0.0702
calm	0.025	0.012	0.035	0.0702
delighted	0.024	0.013	0.074	0.111
attentive	0.022	0.015	0.148	0.19
interested	0.018	0.014	0.189	0.213
relaxed	0.006	0.012	0.617	0.617

The results of the ABCD P&T models generally reflected those of the LDpred scores (Tables S2 and S3). In the white subsample, being interested and able to concentrate were nominally significant. In the P&T scores, being interested was significant at the $p < 0.05$ threshold and showed the same direction of effect as did all the other interested outcome models. Similarly, with concentration, models were significant at the $p < 0.01$ threshold and higher and all models showed the same direction of effect as the LDpred model. Two differences in the P&T models compared to the LDpred models were that being delighted was significant at the $p < 5*10^{-5}$ threshold ($\beta = 0.039$, S.E. = 0.016, $p = 0.013$) and being confident was significant at the $p < 0.05$ ($\beta = 0.038$, S.E. = 0.015, $p = 0.014$) and $p = 1$ thresholds ($\beta = 0.034$, S.E. = 0.016, $p = 0.027$). All other P&T models in the white subsample did not achieve statistical significance.

In the whole sample, being confident, enthusiastic, energetic, able to concentrate and calm were all nominally significant. In the P&T models, being confident and being enthusiastic were both significant at the $p < 0.05$ threshold (confident: $\beta = 0.028$, S.E. = 0.013, $p = 0.029$. enthusiastic: $\beta = 0.027$, S.E. = 0.014, $p = 0.05$). Being energetic was nominally significant at $p < 0.5$ ($\beta = 0.021$, S.E. = 0.009, $p = 0.017$) and $p = 1$ thresholds ($\beta = 0.02$, S.E. = 0.086, $p = 0.02$). Being able to concentrate showed greatest concurrence with all models being nominally significant from $p < 0.01$ and above. This outcome showed the strongest association in the LDpred models and had the most significant thresholds in the P&T models. Being calm was significant at both the $p < 0.5$ ($\beta = 0.024$, S.E. = 0.02, $p = 0.04$) and the $p = 1$ ($\beta = 0.025$, S.E. = 0.02, $p = 0.037$) thresholds. All the other P&T models did not achieve statistical significance. Also note that all P&T models like the LDpred models would not pass FDR correction.

Add Health

A significant association was found in the LDpred analysis in the European ancestry subsample (β = 0.036, S.E. = 0.012, p = 0.0019) and also in the whole sample (β = 0.025, S.E. = 0.009, p = 0.009). The P&T models reflected similar findings with significant associations found in the European ancestry subsample (best threshold p <0.05; β =0.03, S.E. = 0.012, p = 0.01), and in the whole sample (best threshold p <0.1; β =0.022, S.E. = 0.009, p = 0.017) (Tables S4 and S5 respectively). In both the European subsample and in the whole sample the models showed a positive correlation with greater genetic loading for happiness with the exception of the p < 5 $*10^{-5}$ threshold in both groups and p < 0.01 threshold in the whole sample however, none of these models were statistically significant.

MRI

UK Biobank

After correcting for multiple testing via FDR, both (head-size adjusted) white and grey matter volumes (white β = 0.014, S.E. = 0.0053, p.fdr = 0.04; grey β = 0.012, S.E. = 0.0042, p.fdr = 0.016), as well as the general factor for frontal lobe volume (β = 0.019, S.E. = 0.0048, p.fdr = 0.001), remained significant, showing that increasing genetic loading for happiness is associated with higher values of these brain measures (Table S6). We also detected significant associations between genetic loading for general happiness and the left hippocampus tail (β =0.015, S.E. = 0.005, p.fdr = 0.013) and nominally significant associations with the right hippocampus body (β =0.012, S.E. = 0.005, p = 0.021, p.fdr = 0.078) and the left and right thalamus (left: β =0.011, S.E. = 0.0048, p = 0.024, p.fdr = 0.078. right: β =0.01, S.E. = 0.0048, p = 0.038, p.fdr = 0.11).

The DTI MRI parameters MD and FA both associated with genetic loading for happiness. Greater FA (reflecting better white matter integrity) was associated with higher genetic loading for happiness

while the effect was in the opposite direction for MD, for which higher values reflect worse white matter integrity (FA: β = 0.019, S.E. = 0.0056, p.fdr = 0.009; MD: β = -0.021, S.E. = 0.0054, p.fdr < 0.001)

ABCD

In the whole sample (Tables S7) greater right amygdala volume (β = 0.031, S.E. = 0.01, p = 0.005, p.fdr = 0.1), greater white matter volume (β = 0.028, S.E. = 0.01, p = 0.01, p.fdr = 0.1) and greater frontal lobe volume (β = 0.023, S.E. = 0.01, p = 0.035, p.fdr = 0.2) were all nominally significantly associated with increased genetic loading for general happiness, although none passed FDR correction. Both MD and FA, which were significant in UK Biobank, as well as left and right thalamus, which were nominally significant in the UK Biobank, had the same direction of effect in the whole sample ABCD models. The same was true for all the directions of effect in the white sub sample but none of these models were significant (Table S8).

Discussion

We have shown that a single-item, self-reported happiness measure can be used to identify loci in the genome that are associated with happiness, adding to the evidence that individuals do have a baseline happiness level. Given that our sample size was moderate, the heritability estimate was far below that determined by the twin study meta-analysis of Bartels[6]. The mechanism by which the genes at these loci may be affecting the phenotype is unclear; each locus identified in the GWAS does contain genes thought to play a role in neurodevelopment, although this may reflect pleiotropy. The major histocompatibility complex (MHC) region on chromosome 6 contains hundreds of genes in very high LD. Genes in the MHC region have been shown to be involved in several aspects of neural development, including neurite outgrowth and synapse formation[12]. The most statistically significant SNP within the locus identified on chromosome 9 was rs72743916, located in an intron of *ERCC6L2*. Mutations in

this gene have been shown to cause bone marrow failure syndrome 2, symptoms of which include developmental delay and learning difficulties[13]. The third locus, identified on chromosome 20, is found between *EYA2* and *ZMYND8*. No mutations in *EYA2* have been associated with Mendelian disorders, but over-expression has been found in astrocytoma (a type of brain tumour)[14]. ZMYND8 is a transcriptional regulator[15] and has been shown to bind Drebrin which regulates the distribution of ZMYND8 protein between nucleus and the synapse[16].

There were significant genetic correlations between the happiness phenotype and three of the 'big five' personality traits tested and the direction of these effects is generally in agreement with the psychological trait literature. Our results were consistent with a meta-analysis of phenotypic happiness and the big five by Steel et al.[17] as well as work by Hayes and Joseph[18] showing that conscientiousness and extraversion positively correlate with general happiness whereas neuroticism shows a strong negative correlation. The genetic correlation with openness matched that of Hayes and Joseph in that it was negative (though non-significant), whereas Steel et al. showed a significant positive correlation. With agreeableness our results reflected that of Steel et al in that the correlation was positive (though non-significant) whereas Hayes & Joseph showed a non-significant negative correlation.

The results of the LDpred PRS analyses in ABCD and Add Health showed that the genetic loading of this happiness measure not only remained consistent across the lifespan from age 12 to 73 in those of European ancestry but also across multiple ancestral backgrounds giving evidence that the pathways involved are common to a range of ancestral backgrounds. Although the models for the ABCD cohort were not significant there was a consistent direction of association with all nine of the tested traits in both the white subsample and the multi-

ancestry analyses. These models may have been limited by relatively small sample size and also in that the measures used reflect the responses of the parent and not of the child themselves. They are also measures of well-being and not a direct happiness measure.

The use of a novel method involving use of UK Biobank participants as the discovery set for the construction of LDpred scores was also validated by similar results in the P&T PRS that used participants of the respective cohorts to establish LD. This allowed for the sample size to be maximised in these smaller cohorts increasing the statistical power of the analyses.

The results of the MRI analyses detected several brain regions associated with genetic loading for happiness. The frontal lobe has been implicated in hedonic emotions[19,20] and we showed that greater values of a general factor for frontal lobe volume significantly correlated with genetic loading for happiness; however we were unable to replicate this finding in the ABCD cohort which may be due to a lack of power as the sample size was only approximately 7000. It is also the case that the frontal lobes continue to develop until around age 25 years so ABCD participants may yet be too young for this relationship to be demonstrated. We also detected a positive correlation with white and grey matter volumes in the UK Biobank, which was reflected in the ABCD analyses but only white matter in the whole sample achieved nominal statistical significance.

Larger left hippocampal tail and right hippocampal body both associated with greater genetic loading for happiness whereas the whole hippocampal volume was not significantly associated in either UK Biobank or ABCD. The hippocampus outcomes in both cohorts were insignificant which contradicts what has been found elsewhere in the literature, for example in the meta-analysis of Tanzer & Wayandt[21], however, they included parahippocampal

regions along with the hippocampus. There were no data available for sub-hippocampal structures in the ABCD so a direct comparison was not possible.

We also found significant correlations with MD and FA showing higher white matter integrity correlated with greater loading for happiness. This result was reflected in the ABCD cohort but did not reach statistical significance.

Strengths and Limitations

The initial GWAS was large enough to detect significant loci and the use of BOLT-LMM allowed for maximisation of the sample size by accounting for relatedness using the genetic relationship matrix. The GWAS was limited by only including those of white British ancestry. This was due to the imputation panels used in the UK Biobank: both the HRC and the 1000genomes reference panel were European only. Future analyses that include those of other ancestries will be possible once the TOPMED dataset becomes available which has participants imputed using the most appropriate ancestral reference panel. UK Biobank is not representative of the UK general population in that participants are generally healthier and have a higher socioeconomic status than the general population and therefore may have a different happiness level distribution than the UK as a whole[22]. The MRI sample is even less representative in that participation was slightly biased towards the fitter, healthier participants in UK Biobank [23].

A similar issue arises in the PRS analyses in that those of European ancestry were used to establish LD structure of the genome. This was due to the lower numbers of the non-European ancestry participants in these cohorts.

Conclusions

These analyses demonstrate that general happiness level has a genetic contribution which has a consistent effect across age groups and ancestral backgrounds. These analyses not only help increase our understanding of psychology and neurodevelopment, the novel methodology of using UK Biobank participants as a reference panel for LDpred risk scores could be adapted for a wide range of other phenotypes.

It is important to note however, the difference in heritability estimates between our genetic study and those obtained at the phenotypic level suggests that there are still many more genetic loci yet to be discovered that will contribute to baseline happiness level. As such, further analyses will be required using larger datasets with less bias towards those of European ancestry.

Methods

Cohorts, genotyping and phenotyping

UK Biobank

Cohort Description

UK Biobank is a cohort of over half a million UK residents, aged from approximately 40 to 70 years at baseline. It was created to study environmental, lifestyle and genetic factors in middle and older age[8]. Baseline assessments occurred over a 4-year period, from 2006 to 2010, across 22 UK centres. These assessments were comprehensive and included social, cognitive, lifestyle and physical health measures.

UK Biobank obtained informed consent from all participants, and this study was conducted under generic approval from the NHS National Research Ethics Service (approval letter

dated 29 June 2021, Ref 21/NW/0157) and under UK Biobank approvals for application #6553 'Genome-wide association studies of mental health' (PI Rona Strawbridge; GWAS) and #17689 (PI Donald Lyall; imaging).

Genotyping

In March 2018, UK Biobank released genetic data for 487,409 individuals, genotyped using the Affymetrix UK BiLEVE Axiom or the Affymetrix UK Biobank Axiom arrays (Santa Clara, CA, USA) containing over 95% common content. Pre-imputation quality control, imputation and post-imputation cleaning were conducted centrally by UK Biobank (described in the UK Biobank release documentation)[24].

Phenotyping and exclusion criteria

The GWAS was based on the response to the question “In general, how happy are you?” (Data Field 4526), with ordinal responses on a six-point scale ranging from extremely happy to extremely unhappy. Participants were excluded if they were missing more than 10% of their genetic data, if their self-reported sex did not match their genetic sex, if they were determined by UK Biobank to be heterozygosity outliers, and if they were not of white British ancestry (classified by UK Biobank based on self-report and genetic principal components)[24]. Those who had attended for magnetic resonance imaging (MRI) were excluded from the GWAS, in order to maintain independence for subsequent analyses of the MRI phenotypes.

MRI brain scans

Several structural and functional brain MRI measures are available in UK Biobank as imaging derived phenotypes (IDPs)[25]. The brain imaging data, as of January 2021, were used (N=47,920). Brain imaging data used here were processed and quality-checked by UK Biobank and we made use of the IDPs[26,27]. Details of the UK Biobank imaging acquisition and

processing, including structural segmentation and white matter diffusion processing, are freely available from three sources: the UK Biobank protocol: http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=2367 and documentation: http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=1977 and in protocol publications (https://biobank.ctsu.ox.ac.uk/crystal/docs/brain_mri.pdf). Total white matter hyperintensity volumes were calculated on the basis of T1 and T2 fluid-attenuated inversion recovery, derived by UK Biobank. White matter hyperintensity volumes were log-transformed due to a positively skewed distribution. We constructed general factors of white matter tract integrity using principal component analysis. The two separate unrotated factors used were fractional anisotropy (FA), gFA, and mean diffusivity (MD), gMD, previously shown to explain 54% and 58% of variance respectively[28]. We constructed a general factor of frontal lobe grey matter volume using 16 subregional volumes as per Ferguson et al[28]. Total grey matter and white matter volumes were corrected for skull size (by UK Biobank). Models were adjusted for age, sex, genetic principal components (GPCs) 1-8 and the happiness PRS. (see below)

Adolescent Brain Cognitive Development

Cohort description

The Adolescent Brain Cognitive Development (ABCD) cohort is a longitudinal study of brain development and child health[10]. Investigators at 21 sites around the USA conducted repeated assessments of brain maturation in the context of social, emotional, and cognitive development, as well as a variety of health and environmental outcomes. We analysed data from release 3.0. At the time of the survey questions, the children ranged in age from 9 to 12 years. Informed written consent was provided by parents and assent was provided by children, The ABCD research protocol approved by study itself was approved by the Institutional Review Board of University of California San Diego (IRB# 160091) [29].

364 *Genotyping*

365 DNA was extracted from saliva samples of the ABCD participants[30]. These samples were
366 genotyped on the Affymetrix NIDA SmokeScreen Array (Affymetrix, Santa Clara, CA, USA). The
367 QC procedures are described in full at the following URL: https://doi.org/10.15154/1503209.

368 *Phenotyping and exclusion criteria*

369 As no direct measurement of general happiness was available, a set of questions taken from
370 the ABCD Youth NIH Toolbox Positive Affect Items was used instead. These questions
371 measured aspects of positive emotions and affective well-being in the past week, specifically
372 being attentive, delighted, calm, relaxed, enthusiastic, interested, confident, energetic and
373 able to concentrate. Responses were measured as 'not true', 'somewhat true' or 'very true'.
374 Each item was analysed separately.

375 As the initial UK Biobank GWAS was run in the white British sub-group, testing was performed
376 firstly in the white (as defined by ABCD) participants and secondly in the whole sample, with
377 ancestry treated as a factor variable. Participant ancestry was derived from a series of yes/no
378 questions where the respondent's parent selected from a list of options. Those options were:
379 white, black, Native American, Alaskan, Hawaiian, Guamanian, Samoan, Other Pacific
380 Islander, Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian, Other Race,
381 Refuse To Answer, and Don't Know. Those who responded with either of the last two options
382 were excluded. An additional question, "Is your child Hispanic?", was also asked and those
383 who responded yes were defined as Hispanic irrespective of whether they also selected
384 another option. Other categories were then defined as follows: White (n = 4,655), Black (n =
385 1,016), Hispanic (n = 1,201), Native American (Native American and Alaskan, n = 37), East
386 Asian (Chinese, Filipino, Japanese, Korean, Vietnamese and Other Asian, n = 94), Indian (n =

37), Pacific Islanders (Hawaiian, Guamanian, Samoan, Other Pacific Islander, n = 8) and Other Race (n= 39).

MRI

Creation of the derived MRI variables from the ABCD cohort has been described in detail elsewhere[31]. For the purposes of this study, total frontal lobe volume was derived by summing the 22 frontal lobe subsection variables of the left and right hemisphere[32]. Additionally, we looked at total grey and white matter volume and left and right hippocampus volume. The hippocampal body and tail regions and white matter hyperintensity volume were not available for replication. All outcomes were transformed into z scores and all models were adjusted for age, sex, PGCs 1-8, MRI site, and the happiness PRS. For models that included participants from different ancestries, a factor variable for ancestry was included. Models were weighted to match the American community survey (ACS) data by the weighting variable "acs raked propensity score". Relationship filtering was also performed removing one individual at random from any pair of participants who were 2nd cousins or closer with valid phenotypes.

Add Health

Cohort description

Add Health is a nationally representative cohort study of more than 20 000 adolescents from the USA who were aged 12–19 years at baseline assessment in 1994–95. They have been followed through adolescence and into adulthood with five in-home interviews in five waves (I-V) conducted in 1995, 1996, 2001–02, 2008–09 and 2016–18. In this analysis, participants ranged from 24.3 – 34.7 years old, 53% were female and 62% were non-Hispanic white. The study was approved by the University of California San Diego Institutional Review Board (IRB #190002XX).

Genotyping

Saliva samples were obtained as part of the Wave IV data collection. Two Illumina arrays were used for genotyping, with approximately 80% of the sample genotyped with the Illumina Omni1-Quad BeadChip and the remainder of the group genotyped with the Illumina Omni2.5-Quad BeadChip. After quality control, genotyped data were available for 9,974 individuals (7,917 from the Omni1 chip and 2,057 from the Omni2 chip) on 609,130 SNPs present on both genotyping arrays[33]. Imputation was performed separately for European ancestry (imputed using the HRC reference panel) and non-European ancestry samples (imputed using the 1000 Genomes Phase 3 reference panel)[34].For more information on the genotyping and quality control procedures see the Add Health GWAS QC report online at: https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/AH_GWAS_QC.pdf

Phenotyping and exclusion criteria

The outcome variable was collected during the at-home interview of Wave IV and was derived from the response to the question: "How often was the following true during the past seven days? You felt happy." Responses were given as: "never or rarely"; "sometimes"; "a lot of the time"; "most of the time or all of the time"; "refused"; "don't know". Those who responded with the latter two options were excluded. Remaining categories were coded from "never" = 0 to "all of the time" = 3.

Ancestry in Add Health is defined in the 'psancest' variable as European, African, Hispanic and East Asian. Additionally, Add Health provides a weighting variable to make the results reflective of the US population. In these analyses the models were weighted by the Wave IV variable 'gswgt4_2'. All p values for PRS analyses were False Discovery Rate (FDR)-adjusted[35].

Analyses

Genetic Association

Genetic association analysis was performed in UK Biobank using BOLT-LMM[9,36] treating the outcome variable as a quasi-quantitative trait. BOLT-LMM uses a genetic relationship model constructed from genotyped SNPs, which allows for maximisation of sample size by adjusting for relatedness, thus also avoiding the need to adjust the model for GPCs. The model was adjusted for age, sex and genotyping array. SNPs were filtered by minor allele frequency (MAF) > 0.01, Hardy-Weinberg equilibrium $p > 1 \times 10^{-6}$, and imputation quality score > 0.8, leaving 9,286,407 SNPs for testing. BOLT-REML[37] was also used to provide a heritability estimate and λ_{GC} estimate to test for inflation of the test statistics.

Linkage disequilibrium score regression

Linkage disequilibrium (LD) score regression was performed using LDSC[38]. The summary statistics from the GWAS carried out using the UK Biobank data were compared against four of the 'big five' personality traits: openness, conscientiousness, extraversion and agreeableness from Lo et al[39]. For neuroticism we used the summary statistics from a GWAS reported in Smith et al[40] due to the larger sample size and a greater number of significant loci found compared to Lo et al's[39] neuroticism output.

LDpred risk score generation

UK Biobank

LDpred[41] established the LD structure of the genome using a reference panel of 1000 unrelated white British UK Biobank participants (the PRS discovery set). These participants had been excluded from the GWAS, due to having a missing phenotype or because they had valid MRI data, but still passed the same QC as described above. SNPs were filtered using the same parameters as for the GWAS. Scores were then created in the validation set using an

455 infinitesimal model. Models using polygenic risk scores (PRS) derived using LDpred were
456 adjusted for age, sex, genotyping array and the first eight GPCs.

457 *ABCD and Add Health*

458 Due to the lower cohort size of ABCD and Add Health, it would not have been possible to
459 remove 1000 participants from the analyses to use as a discovery set without markedly
460 reducing the power of the analyses. Therefore, we used the 1000 unrelated UK Biobank
461 participants as the discovery set to establish LD and this was used to generate the risk scores
462 for the participants in these datasets[42]. The only additional step was to find the SNPs that
463 were found in both the discovery and validation datasets and passed the same SNP filtering
464 criteria in both datasets, with an additional filter that MAF threshold was set at > 0.05 due to
465 the lower mean imputation quality of the less common SNPs found in these smaller cohorts.
466 Models were additionally adjusted for age at view, sex and the first 10 GPCs. For multi-
467 ancestry models, ancestry was treated as a factor variable.

468 Pruning and Threshold PRS generation in ABCD and Add Health

469 To ensure that using the UK Biobank participants as a reference panel was not introducing
470 biases into the LDpred risk scores in the ABCD and Add Health cohorts, pruning and
471 thresholding (P&T) PRSs were also created using participants from within the respective
472 cohorts to corroborate the findings of the LDpred PRS models. Linkage disequilibrium was
473 established using a reference of unrelated white participants from each their respective
474 cohorts, as described above. SNPs were clumped in plink[43] using a cut-off of $r^2 > 0.1$ in a 500kb
475 window. SNPs were filtered using p-value thresholds of $p < 5 \times 10^{-5}$, $p < 0.01$, $p < 0.05$, $p < 0.1$,
476 $p < 0.5$ and $p < 1$.

References

1 Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine* **11**, 126, doi:10.1186/1741-7015-11-126 (2013).

2 Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 624-633, doi:10.1038/ng.3552 (2016).

3 Mancini, A. D., Bonanno, G. A. & Clark, A. E. Stepping Off the Hedonic Treadmill. *Journal of Individual Differences* **32**, 144-152, doi:10.1027/1614-0001/a000047 (2011).

4 Brickman, P., Coates, D. & Janoff-Bulman, R. Lottery winners and accident victims: is happiness relative? *J Pers Soc Psychol* **36**, 917-927, doi:10.1037//0022-3514.36.8.917 (1978).

5 Nes, R. B. & Røysamb, E. Happiness in Behaviour Genetics: An Update on Heritability and Changeability. *Journal of Happiness Studies* **18**, 1533-1552, doi:10.1007/s10902-016-9781-6 (2017).

6 Bartels, M. Genetics of Wellbeing and Its Components Satisfaction with Life, Happiness, and Quality of Life: A Review and Meta-analysis of Heritability Studies. *Behavior Genetics* **45**, 137-156, doi:10.1007/s10519-015-9713-y (2015).

7 Shu, L., Blencowe, M. & Yang, X. Translating GWAS Findings to Novel Therapeutic Targets for Coronary Artery Disease. *Front Cardiovasc Med* **5**, 56-56, doi:10.3389/fcvm.2018.00056 (2018).

8 Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

9 Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-290, doi:10.1038/ng.3190 (2015).

10 Jernigan, T. L., Brown, S. A. & Dowling, G. J. The Adolescent Brain Cognitive Development Study. *J Res Adolesc* **28**, 154-156, doi:10.1111/jora.12374 (2018).

11 Harris, K. M. The National Longitudinal Study of Adolescent to Adult Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002; Wave IV, 2007-2009 [machine-readable data file and documentation]. *Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.* (2009).

12 McAllister, A. K. Major Histocompatibility Complex I in Brain Development and Schizophrenia. *Biological Psychiatry* **75**, 262-268, doi:10.1016/j.biopsych.2013.10.003 (2014).

13 Shabanova, I. *et al.* ERCC6L2-associated inherited bone marrow failure syndrome. *Mol Genet Genomic Med* **6**, 463-468, doi:10.1002/mgg3.388 (2018).

14 Wen, Z., Liang, C., Pan, Q. & Wang, Y. Eya2 overexpression promotes the invasion of human astrocytoma through the regulation of ERK/MMP9 signaling. *Int J Mol Med* **40**, 1315-1322, doi:10.3892/ijmm.2017.3132 (2017).

15 Malovannaya, A. *et al.* Analysis of the human endogenous coregulator complexome. *Cell* **145**, 787-799, doi:10.1016/j.cell.2011.05.006 (2011).

16 Yao, N. *et al.* The Structure of the ZMYND8/Drebrin Complex Suggests a Cytoplasmic Sequestering Mechanism of ZMYND8 by Drebrin. *Structure* **25**, 1657-1666.e1653, doi:https://doi.org/10.1016/j.str.2017.08.014 (2017).

17 Steel, P., Schmidt, J. & Shultz, J. Refining the Relationship Between Personality and Subjective Well-Being. *Psychological bulletin* **134**, 138-161, doi:10.1037/0033-2909.134.1.138 (2008).

18 Hayes, N. & Joseph, S. Big 5 correlates of three measures of subjective well-being. *Personality and Individual Differences* **34**, 723-727, doi:https://doi.org/10.1016/S0191-8869(02)00057-0 (2003).

19 Kringelbach, M. L. & Berridge, K. C. The Neuroscience of Happiness and Pleasure. *Soc Res (New York)* **77**, 659-678 (2010).

20 Tăbăcaru, B. in *2013 E-Health and Bioengineering Conference (EHB).* 1-4.

21 Tanzer, J. R. & Weyandt, L. Imaging Happiness: Meta Analysis and Review. *Journal of Happiness Studies* **21**, 2693-2734, doi:10.1007/s10902-019-00195-7 (2020).

22 Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American journal of epidemiology* **186**, 1026-1034, doi:10.1093/aje/kwx246 (2017).

23 Lyall, D. *et al. Quantifying bias in psychological and physical health in the UK Biobank imaging sub-sample*. (2021).

24 Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298, doi:10.1101/166298 (2017).

25 Elliott, L. T. *et al.* The genetic basis of human brain structure and function: 1,262 genome-wide associations found from 3,144 GWAS of multimodal brain imaging phenotypes from 9,707 UK Biobank participants. *bioRxiv*, 178806, doi:10.1101/178806 (2017).

26 Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* **19**, 1523-1536, doi:10.1038/nn.4393 (2016).

27 Smith, S. M. *et al.* Enhanced Brain Imaging Genetics in UK Biobank. *bioRxiv*, 2020.2007.2027.223545, doi:10.1101/2020.07.27.223545 (2020).

28 Ferguson, A. C. *et al.* Association of SBP and BMI with cognitive and structural brain phenotypes in UK Biobank. *J Hypertens* **38**, 2482-2489, doi:10.1097/hjh.0000000000002579 (2020).

29 Auchter, A. M. *et al.* A description of the ABCD organizational structure and communication framework. *Developmental Cognitive Neuroscience* **32**, 8-15, doi:https://doi.org/10.1016/j.dcn.2018.04.003 (2018).

30 Uban, K. A. *et al.* Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data. *Developmental Cognitive Neuroscience* **32**, 97-106, doi:https://doi.org/10.1016/j.dcn.2018.03.005 (2018).

31 Hagler, D. J., Jr. *et al.* Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage* **202**, 116091, doi:10.1016/j.neuroimage.2019.116091 (2019).

32 Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968-980, doi:10.1016/j.neuroimage.2006.01.021 (2006).

33 (Highland, H. M. A., Christy L.; Duan, Qing; Li, Yun; Mullan Harris, Kathleen 2018). *Polygenic Scores (PGSs) in the National Longitudinal Study of Adolescent to Adult Health (Add Health) – Release 2*.

34 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

35 Benjamini, Y. & Hochberg, Y. *Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing*. Vol. 57 (1995).

36 Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* **50**, 906-908, doi:10.1038/s41588-018-0144-6 (2018).

37 Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nature Genetics* **47**, 1385-1392, doi:10.1038/ng.3431 (2015).

38 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295, doi:10.1038/ng.3211 http://www.nature.com/ng/journal/v47/n3/abs/ng.3211.html#supplementary-information (2015).

39 Lo, M.-T. *et al.* Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nature genetics* **49**, 152-156, doi:10.1038/ng.3736 (2017).

40 Smith, D. J. *et al.* Genome-wide analysis of over 106000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry* **21**, 749-757, doi:10.1038/mp.2016.49 (2016).

41 Vilhjálmsson, Bjarni J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics* **97**, 576-592, doi:10.1016/j.ajhg.2015.09.001 (2015).

42 Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLOS Genetics* **17**, e1009021, doi:10.1371/journal.pgen.1009021 (2021).

43 Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* **81**, 559-575 (2007).