Independent Polygenic Component Scores of Brain Structure and Function

Lennart M. Oblong*[1], Nicolò Trevisan[1], Sourena Soheili-Nezhad[1], Yingjie Shi[2], Christian F. Beckmann[1,3], Emma Sprooten[1,2]

[1]Department of Medical Neuroscience, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Centre, Nijmegen, The Netherlands

[2]Department of Human Genetics, Donders Institute for Brain, Cognition and Behavior, Radboud University Medical Centre, Nijmegen, The Netherlands

[3]Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, The Netherlands

***Corresponding author:** Lennart M. Oblong

lennart.oblong@donders.ru.nl

Department of Medical Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Centre, Nijmegen, The Netherlands

Key words
Genomic ICA, Independent Component Analysis, MELODIC, neuroimaging, GWAS, Multivariate Genomics, Statistical Genetics, Neuroimaging Genetics, Polygenic Scores

Abstract

The polygenic and pleiotropic nature of behavioral, mental health and brain phenotypes complicates mechanistic interpretation of related genome-wide association study (GWAS) signals. We previously proposed genomic independent component analysis (genomICA) as a multivariate method to decompose a large set of univariate GWAS statistics of multimodal brain traits into statistically independent and potentially more interpretable genomic components. Here, we assess the out-of-sample predictive power and stratification potential of this method using polygenic scores derived from genomICA components. We posit that these neuroimaging-based polygenic component scores (PCS) capture pleiotropic polygenic patterns in a structured and interpretable way and explain phenotypic variance reliably, while simultaneously decorrelating individual genetic profiles to reveal independent dimensions of polygenic influence across brain imaging and other traits.

We used the SNP-loadings of our previously derived genomICA components, based on thousands of brain traits, as weights to compute PCS in an independent sample. We first assessed collinearity among PCS to confirm independence. Then, we tested the PCS predictive power in 1269 neuroimaging phenotypes and 858 non-neuroimaging phenotypes using general linear and logistic regression models, and quantified phenotypic variance jointly explained by all PCS. We then examined PCS correlations across individuals to assess whether they offer utility for stratifying phenotypic variation.

PCSs showed little collinearity (r_{min} = -0.04, r_{mean} = $4*10^{-4}$, r_{max} = 0.05). They explained substantial variance in brain traits out-of-sample (R^2_{max}=0.12, R^2_{mean}=0.04). Variance explained in non-neuroimaging phenotypes was much lower (R^2_{max} =0.02, R^2_{mean} = $2*10^{-4}$) but significant. Different individual PCS were associated with distinct groups of neuroimaging categories and brain tissues. PCSs explained meaningful patterns of phenotypic variance in non-neuroimaging phenotypes, showing a distinction along different categories of phenotypes. For instance, PCS 8 clearly captured associations with lifestyle, behavior, diet, and socioeconomic factors, while PCS 15 was linked to a wide range of cardiovascular health outcomes in both participants and their relatives.

These results demonstrate that genomICA-derived PCSs capture independent, uncorrelated patterns of associations across the brain and can predict substantial variance in a range of neuroimaging and other out-of-sample phenotypes. PCSs capture substantial variance in neuroimaging phenotypes and, to a lesser extent, non-neuroimaging traits, with individual PCS aligned to distinct phenotype domains. The inclusion of the brain phenotypes, but not other phenotypes, in the input for the decomposition likely accounts for this difference. GenomICA identifies mutually-independent dimensions in the space of genomic effects, which has utility for stratification of heterogeneous samples. Notably, the dominant genomic influences on the brain

captured here reflect physical traits, lifestyle and environment but not mental health or neurological traits. Overall, this study highlights the potential of genomICA-based PCSs in modelling phenotypic variance across data domains. Future work will focus on refining these components to better capture variance in complex neuropsychiatric and behavioural traits by incorporating them in the decomposition process.

Introduction

Most individual differences across brain, behaviour, physical attributes, physical health and mental health are heritable[1–11] and shaped by highly pleiotropic and polygenic effects[10,12–15] . The inherent complexity of causal biological pathways that subserve complex traits has limited the mechanistic interpretation of genome-wide association studies (GWAS). Individual effects of genome-wide significant single nucleotide polymorphisms (SNPs) are weak, even after aggregation across hundreds or thousands of SNPs, explain a relatively small portion of trait variance ($0.019 < R^2 < 0.147$)[16–22]. The causal chain from SNP to trait is very long and spans multiple biological processes including gene expression, protein synthesis, DNA methylation, cellular processes, 'house-keeping' mechanisms, and environmental interaction. A SNP may exert impact on one or multiple levels of this causal chain, and show associations with a wide range of traits (i.e. pleiotropy), which makes it difficult to interpret the GWAS signal in terms of causal biological processes [10,13,15,23].

The polygenicity and pleiotropy propagates to down-stream analyses such as polygenic score (PGS) analyses, where the PGS aggregates the many biological processes underlying genomic associations into a single score. PGS analysis is widely used to estimate individual differences in genetic propensity for a trait based on individuals' genotypes. In standard PGS calculation, individuals' allele frequency dosages are summed up, and weighted by GWAS SNP effect sizes[24–28]. The power of any PGS to stratify study populations based on disease risk, or (etiological) subtypes is potentially complicated by the unresolved mixture of polygenic pleiotropic genetic effects that are tangled together into a single PGS. Pleiotropic SNPs impact PGS-based stratification due to their lack of trait specificity, introduce collider bias, and create spurious associations driven by genetic correlation. Arguably, standard PGSs generated from a case/control GWAS cannot distinguish phenotype subtypes, as individuals predisposed to one trait/subtype may also have a propensity for another, leading to less precise stratification. PGS-based stratification that takes pleiotropy into account is therefore desirable since it can uncover more specific propensity profiles, and their subserving genetic correlates[29].

Multiple multivariate methods have been proposed that leverage pleiotropic effects using GWAS data. These include genomic structural equation modeling (SEM)[30], genomic Principal Component Analysis (PCA)[31], the Multivariate Omnibus Statistical Test (MOSTest)[32], and Genetic Unmixing by Independent Decomposition (GUIDE)[33]. Genomic SEM requires manual specification of model order (i.e. number of latent factors) and the mapping of traits onto latent factors. This constrains the analysis to a relatively small number of prespecified latent factors. MOSTest combines SNP statistics across traits in a single genome-wide signature, which limits the method's power to recover multiple distinct genomic patterns of pleiotropic effects. Genomic PCA on the other hand is entirely data-driven, but relies on a PCA algorithm which is designed to maximize

the overall explained variance by components without optimizing for mutual independence. Therefore, multiple dominant biological signals that are correlated may be aggregated into a single component. A data-driven approach that decomposes high-dimensional datasets into statistically *independent* components offers a promising avenue to unravel the complex genetic architecture underlying phenotypes[33,34]. We previously introduced genomic independent component analysis (genomICA) to extract maximally independent hidden sources of genomic signal from neuroimaging trait GWAS results, demonstrating the reproducibility of the genomICA latent components[34].

We can view GWAS output as the final endpoints of a 'mixture' of biological processes, where groups of SNPs may share involvement in a pathway, reflected in the covariance of their effect sizes across traits. Multivariate independent component analysis (ICA) is well suited to unmix these processes. As an unsupervised source separation method, ICA decomposes complex signal into maximally independent components, assuming a linear combination of (non-Gaussian distributed) sources and structured noise. Unlike PCA, ICA maximises independence, making it more effective for recovering distinct sources from noisy data[35]. GenomICA[34] assumes GWAS summary statistics to be a linearly mixed signal containing structured and Gaussian noise, thereby leveraging polygenicity and pleiotropy in a fully data-driven framework.

Our previous work has shown that genomICA yields genomic principal and independent components that exceed raw GWAS reproducibility and group genetic variants according to their similarity of associations across different brain features detected through magnetic resonance imaging (MRI)[34]. The shared genetic architecture among neuroimaging derived brain phenotypes captured in these components may reflect patterns of pleiotropy that are not limited to brain traits alone. Since many genetic influences on the brain are correlated with various non-neuroimaging traits[36,37], components derived from the shared variance between brain phenotypes may capture patterns that associate with non-neuroimaging phenotypes, providing a structured and interpretable basis for stratifying heterogeneity. Each component is made of two vectors, one containing the SNP-loadings and the other containing the trait-loadings. The SNP-loadings constitute the independent genomic sources of genetic variance, and can be used as a more differentiated "pseudo-GWAS summary statistic" that is linked to a specific combination of brain-phenotypes. This is achieved by combining multiple correlated trait associations into a single SNP-loading. Simultaneously, the components disentangle polygenic effects by separating genome-wide SNP effects into distinct PGSs, based on the independence of their associations across many traits. These genome-wide SNP-loadings offer a new approach to aggregating polygenic effects in the form of a polygenic component score (PCS), the "unmixed" independent nature of which could improve the potential of PGS methods to stratify target populations.

In the present exploratory study we used the genomic components in a polygenic scoring framework as the base GWAS-like sample to calculate PCSs for each out-of-sample participant. Thereby, we seek to test the potential of genomICA derived components to improve stratification and PGS specificity across a wide range of phenotypes. By leveraging pleiotropy across thousands of brain-GWAS imaging phenotypes, this method is designed to reconstruct mutually independent genetic patterns associated with brain features, which in turn may improve PGS to achieve more specific stratification of study populations. First, we tested out-of-sample PCS's predictive power for the same neuroimaging phenotypes that were used in component creation. Then, we tested component PCSs on hundreds of non-neuroimaging phenotypes that include general physical attributes, nutritional questionnaire data, physical health, mental health, general well-being, lifestyle, behaviour, socio-economic, and demographic metrics. GenomICA decomposition may increase the sensitivity of polygenic prediction, improve our understanding of the biological pathways subserving trait development, differences in the relevance of these pathways, and simultaneously generate insights into interlinking environmental influences.

Results

Polygenic Component Scores of multivariate genomic independent and principal components are orthogonal

Pearson-correlation between all PCSs showed that PCSs derived respectively from genomic ICs and PCs are largely uncorrelated, which serves as validation that the genomic patterns subserving their calculation are likewise independent at the individual-level in an independent sample (PGS-PC1 from genomic IC 1-16: r_{min} = -0.04, r_{mean} = $4*10^{-4}$, r_{max} = 0.05; PGS-PC 1 from genomic PC1-16: r_{min} = -0.04, r_{mean} = $5*10^{-4}$, r_{max} = 0.04)(figure 2). Correlations between the IC-PCSs and PC-PCSs were high, which is expected given that the ICs are a rotation of the PCs and both sets of components capture variance in the same data space. Given this low collinearity, we performed all subsequent analyses using models containing all 16 PCSs simultaneously as predictors for each trait.

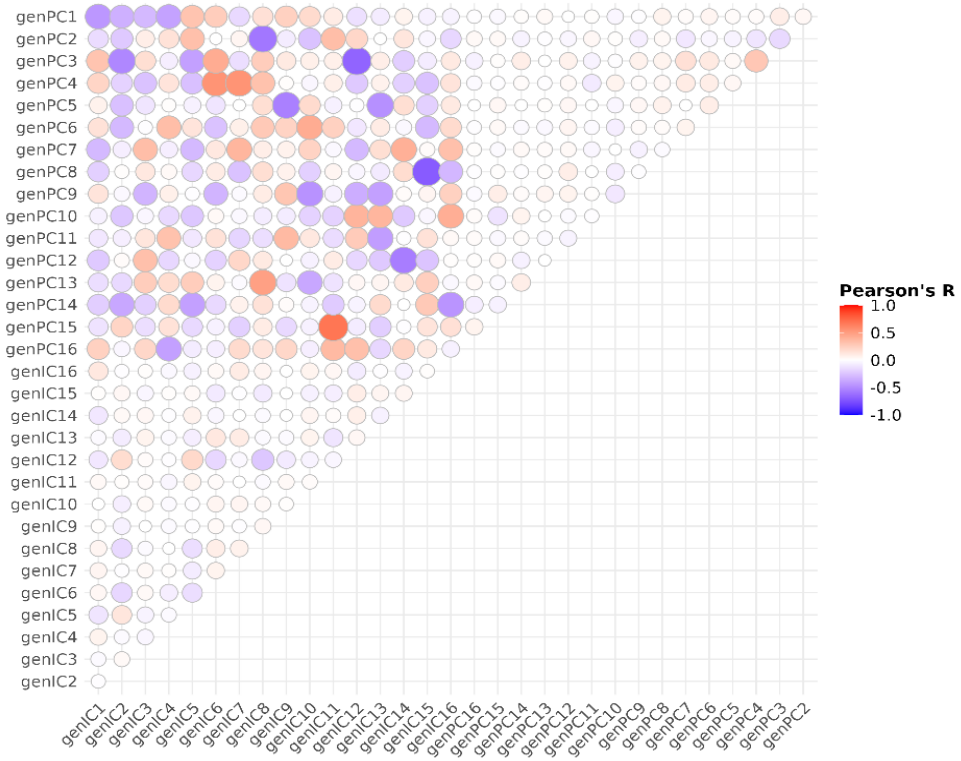

Figure 2: Correlation heatmap resulting from the correlation of the first polygenic score PC derived from all 16 genomic IC and genomic PC. The upper-left 16-by-16 square indicated the correlations of genomic PC derived polygenic PCs with the genomic IC derived polygenic PCs. The upper-right and lower-left triangles indicate the correlations between genomic PC derived polygenic PCs and genomic IC derived polygenic PCs, respectively. IC – independent component; PC – principal component; PRS - polygenic risk score

Neuroimaging-derived polygenic component scores are associations with neuroimaging phenotypes in an independent sample.

We first assessed the performance of our neuroimaging-derived PCSs on the same brain imaging phenotypes that were included in the GenomICA pipeline, in the independent target sample. GenomICA derived PCSs captured statistically significant variance in N = 1256 phenotypes post-FDR ($p_{FDR} < 0.05$). Polygenic variance explained varied across neuroimaging categories (figure 3). Neuroimaging QC metrics ($R^2_{mean} = 0.09$), dMRI TBSS metrics ($R^2_{mean} = 0.07$), and dMRI ProbtrackX metrics ($R^2_{mean} = 0.07$) show the highest average phenotypic variance explained. For individual neuroimaging traits, the highest (conventional) polygenic variance was captured in global brain features such as *T1-imaging scan to standard-space warping*, a proxy for intracranial volume (ICV), large white-matter tract *orientational discrepancy (OD) of the posterior limb of the internal capsule* in both hemispheres, and *global normalized cerebrospinal fluid (CSF) volume* derived from T1-weighted SIENAX tissue segmentation (table 2). On the whole, the highest R^2-values are achieved in metrics related to large white matter tracts with half of the top 20 IDPs related to isotropic diffusion values which generally reflect tissue density (table 2).

As a further confirmation of the out-of-sample reproducibility of the method, we found that PCS effect sizes on neuroimaging phenotypes in the target sample were highly correlated with the neuroimaging trait-loadings in the discovery sample ($0.54 < r < 0.92$; table S3).

Since PCS-R^2 is bounded by SNP-h^2, we also calculated the $R^2_{conventional}$ as a fraction of the SNP-heritability (h^2_{SNP}). For some phenotypes, such as the *T1-weighted image to standard-space warping* PCSs collectively explain a large fraction of the heritability ($R^2_{SNP} / h^2_{SNP} = 0.82$). Across all neuroimaging phenotypes, the average h^2-fraction explained by PCS was 0.17. Similar results for other neuroimaging phenotypes are presented in full in table S4 and figure S1.

The results for the genomic PCA derived PCS analysis were similar in R^2-magnitude as those for the genomic ICA PCS, and can be found in the supplement (figure S2 & S3; table S5).

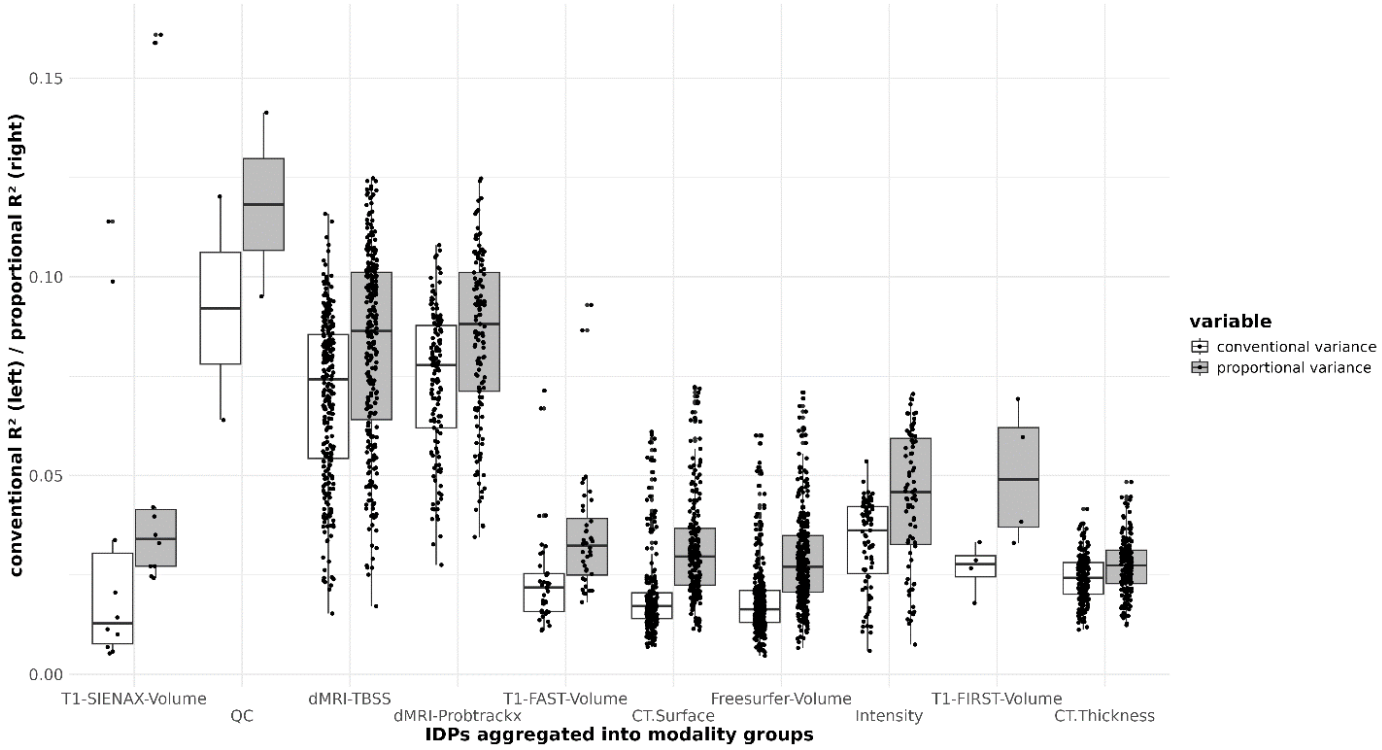

Figure 3: Overview of neuroimaging phenotypic variance explained by genomic IC PCSs significant after FDR-correction and aggregated into categories, indicated on the x-axis. Shown is the conventional variance explained (white), and proportional variance explained (grey). The box-plot central line indicates the mean within the category and the outer borders the 1st and 3rd quartiles. CT – cortical; IC – independent component; IDP – imaging derived phenotype; ProbtrackX – probabilistic tractography; TBSS – tract-based spatial statistics.

Table 1: Table showing the top 20 IDPs predicted by PCSs in concert, sorted by Nagelkerke's pseudo-R^2. P-values are adjusted with FDR-correction.

ID	*Description*	*Full.model.R^2*	*Null.model.R^2*	*PCS.R^2*	*PCS.R^2.adj*	*p.value*	*Modality*
25733	T1-to-standard_nonlinear_alignment_warping	0.27	0.149	0.12	0.141	2.54e-72	QC
25411	OD_Posterior_limb_of_internal_capsule_L	0.168	0.0517	0.116	0.122	1.28e-60	dMRI-TBSS
25003	CSF_normalised_volume	0.406	0.292	0.114	0.161	1.55e-84	T1-SIENAX-Volume
25410	OD_Posterior_limb_of_internal_capsule_R	0.169	0.0556	0.114	0.121	7.02e-60	dMRI-TBSS
25365	ICVF_Retrolenticular_part_of_internal_capsule_L	0.207	0.097	0.11	0.122	1.85e-60	dMRI-TBSS
25481	ISOVF_Superior_longitudinal_fasciculus_L	0.213	0.105	0.108	0.121	7.02e-60	dMRI-TBSS
25671	ICVF_slf_l	0.238	0.13	0.108	0.124	1.02e-61	dMRI-Probtrackx
25581	L1_ifo_l	0.216	0.11	0.107	0.12	1.96e-59	dMRI-Probtrackx
25337	L3_Superior_longitudinal_fasciculus_L	0.248	0.142	0.107	0.124	1.02e-61	dMRI-TBSS
25662	ICVF_ifo_l	0.221	0.115	0.105	0.119	4.52e-59	dMRI-Probtrackx
25725	ISOVF_slf_l	0.207	0.102	0.105	0.117	1.09e-57	dMRI-Probtrackx
25408	OD_Anterior_limb_of_internal_capsule_R	0.162	0.0579	0.104	0.11	4.52e-54	dMRI-TBSS
25385	ICVF_Superior_longitudinal_fasciculus_L	0.23	0.127	0.103	0.118	1.93e-58	dMRI-TBSS
25672	ICVF_slf_r	0.223	0.121	0.102	0.116	1.87e-57	dMRI-Probtrackx
25500	FA_ifo_l	0.182	0.0799	0.102	0.111	2.71e-54	dMRI-Probtrackx
25527	MD_ifo_l	0.229	0.128	0.101	0.116	3.69e-57	dMRI-Probtrackx
25417	OD_Superior_corona_radiata_L	0.18	0.079	0.101	0.109	1.97e-53	dMRI-TBSS
25145	MD_Superior_longitudinal_fasciculus_L	0.269	0.169	0.1	0.121	7.02e-60	dMRI-TBSS
25364	ICVF_Retrolenticular_part_of_internal_capsule_R	0.239	0.139	0.1	0.116	2.6e-57	dMRI-TBSS
25730	ISOVF_unc_r	0.162	0.0622	0.0998	0.106	8.8e-52	dMRI-Probtrackx

Neuroimaging-based polygenic component scores associated with non-neuroimaging phenotypes

Overall, many non-neuroimaging phenotypes were highly significantly associated with neuroimaging-derived PCSs, but effect sizes (R^2) were much lower than for the neuroimaging phenotypes themselves.

Looking at the broader domain categories, the largest variance explained were in the domains of *physical health*, *lifestyle and behavior*, and numeric memory in the domain of *cognitive performance* (figure 4).

Individual binary phenotypes most strongly associated with neuroimaging-based PCS were *self-reported Nephritis,* and *self-reported Malabsorption/Coeliac disease,* followed by *frequency of memory loss due to drinking alcohol in the past year* and *frequency of feeling remorse after drinking alcohol in the past year* (table 3). Weaker in effect size, but also with highly significant associations were measures of *physical attributes*, such as the *comparative height at age 10*, and *physical health* measures related to cardiovascular health such as *self-reported high blood-pressure*, *diagnosed high blood pressure*, and *no diagnosis of any cardiovascular health problems*. The full results are shown in the supplement (table S6).

In continuous phenotypes the strongest associations were likewise found in physical health and physical attributes, including height (R^2 = 0.0016, p_{FDR} = 1.9*10^{-232}), weight (R^2 = 4e-4, p_{FDR} = 9.5*10^{-206}), basal metabolic rate (R^2 = 3.9*10^{-4}, p_{FDR} = 1.7*10^{-144}), and systolic (R^2 = 7.9*10^{-4}, p_{FDR} = 2.4*10^{-52}) and diastolic blood-pressure (R^2 = 9.5*10^{-4}, p_{FDR} = 1.1*10^{-60}) (table S8).

We calculated the h^2-fraction explained by PCSs for all tested phenotypes and found lower ratios than those in neuroimaging phenotypes, with an average h^2-fraction of 0.01 (h^2-fraction$_{max}$ = 0.28) in binary and binarized phenotypes (table S6), and 0.004 in continuous phenotypes (h^2-fraction$_{max}$ = 0.02) (table S8).

In the analysis of genomic PCA derived PCSs we find a similar pattern of associations in that the strongest domains of association are those of *physical health*, *lifestyle and behaviour*, *cognitive performance* and *physical attributes*. In the following sections we focus on the results of genomic ICA PCSs considering constraints on manuscript length. The full results of this genomic PCA analysis are presented in the supplement (table S7 & S9; figure S3 & S4).

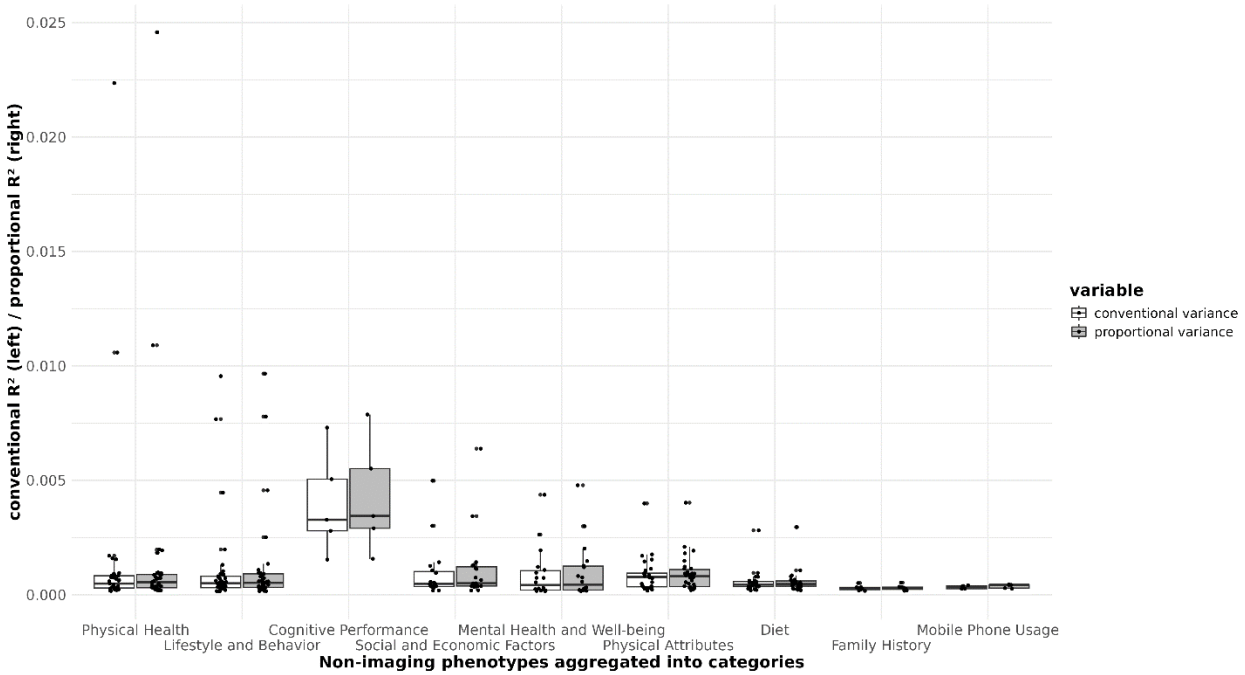

Figure 4: Overview of binary/categorical phenotypic variance explained by genomic IC PCSs significant after FDR-correction and aggregated into domains, indicated on the x-axis. Shown is the conventional Nagelkerke's variance explained (white), and proportional Nagelkerke's variance explained (grey). The box-plot central line indicates the mean within the category and the outer borders the 1st and 3rd quartiles.

Table 2: Table showing the top 20 binary/binarized non-imaging phenotypes predicted by PCSs in concert, sorted by R2. P-values are adjusted with FDR-correction.

Phenotype	*base.Phenotype*	*PGS.Nag.R2*	*full.Nag.R2*	*null.Nag.R2*	*p.value*
self-reported Nephritis	Non-cancer illness code, self-reported	0.022	0.113	0.09	0.014
self-reported Malabsorption/Coeliac disease	Non-cancer illness code, self-reported	0.011	0.04	0.029	1.2e-27
Frequency of memory loss due to drinking alcohol in last year 2 vs 3	Frequency of memory loss due to drinking alcohol in last year	0.0096	0.021	0.012	0.046
Frequency of feeling guilt or remorse after drinking alcohol in last year 2 vs 4	Frequency of feeling guilt or remorse after drinking alcohol in last year	0.0077	0.022	0.015	0.042
Numeric memory test - Maximum digits remembered correctly 5 vs 10	Numeric memory test - Maximum digits remembered correctly	0.0073	0.08	0.073	0.034
Numeric memory test - Maximum digits remembered correctly 5 vs 9	Numeric memory test - Maximum digits remembered correctly	0.0051	0.089	0.084	0.0041
Financial situation satisfaction 1 vs 6	Financial situation satisfaction	0.005	0.224	0.219	0.04
Frequency of memory loss due to drinking alcohol in last year 1 vs 3	Frequency of memory loss due to drinking alcohol in last year	0.0045	0.026	0.021	0.035
Happiness.0 1 vs 4	Happiness	0.0044	0.091	0.087	0.044
Comparative height size at age 10 1 vs 3	Comparative height size at age 10	0.004	0.013	0.0086	9.5e-84
Numeric memory test - Maximum digits remembered correctly 5 vs 8	Numeric memory test - Maximum digits remembered correctly	0.0033	0.051	0.048	0.0021
Financial situation satisfaction 1 vs 4	Financial situation satisfaction	0.003	0.125	0.122	0.0042
Numeric memory test - Maximum digits remembered correctly 6 vs 9	Numeric memory test - Maximum digits remembered correctly	0.0028	0.042	0.039	0.0088
Oily fish intake 0 vs 4	Oily fish intake	0.0028	0.051	0.048	0.018
Happiness and subjective well-being - General happiness with own health 1 vs 4	Happiness and subjective well-being - General happiness with own health	0.0026	0.124	0.121	0.023
Amount of alcohol drunk on a typical drinking day.0 1 vs 5	Amount of alcohol drunk on a typical drinking day	0.002	0.213	0.211	0.043
Traumatic events - Able to pay rent/mortgage as an adult 2 vs 4	Traumatic events - Able to pay rent/mortgage as an adult	0.0019	0.039	0.037	0.022
Skin colour 1 vs 4	Skin colour	0.0018	0.086	0.084	0.047
self-reported Hypertension	Non-cancer illness code, self-reported	0.0017	0.138	0.136	1,00E-62
Relative age of first facial hair male 1 vs 3	Relative age of first facial hair male	0.0017	0.061	0.06	0.027

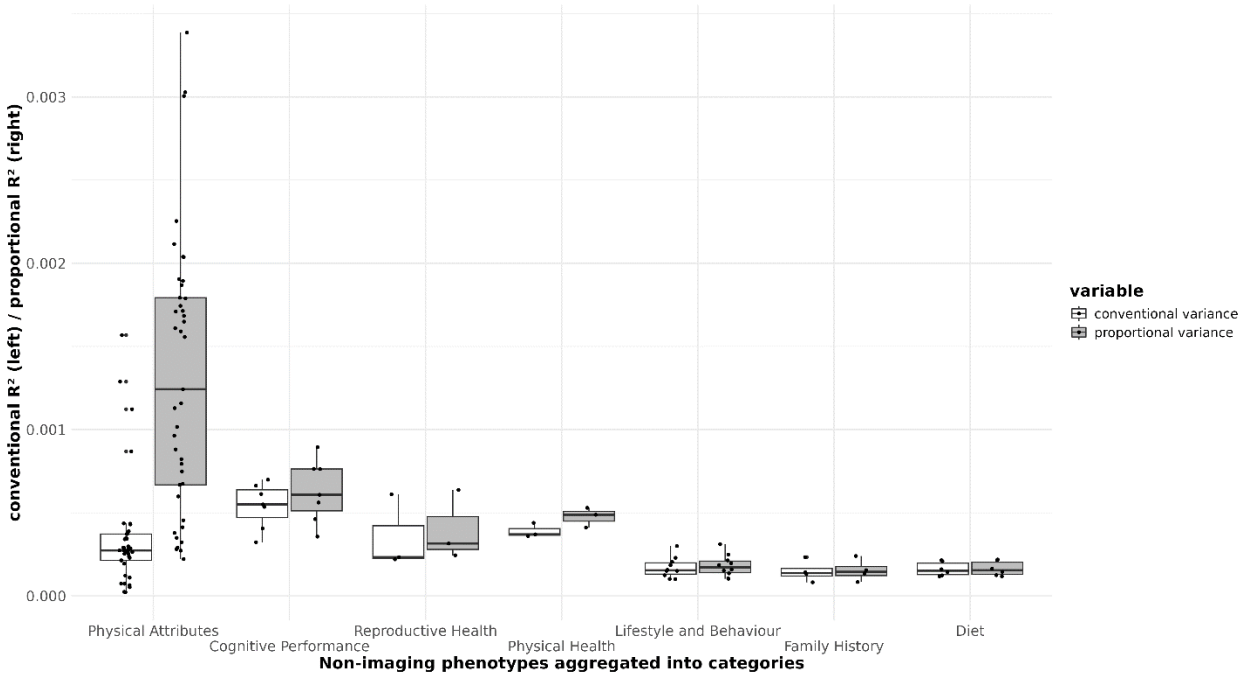

Figure 5: Overview of continuous phenotypic variance explained by genomic IC PCSs significant after FDR-correction and aggregated into domains, indicated on the x-axis. Shown is the conventional Nagelkerke's variance explained (white), and proportional Nagelkerke's variance explained (grey). The box-plot central line indicates the mean within category and the outer borders the 1st and 3rd quantiles.

Overall patterns of phenotypic associations emerge across polygenic component scores

Next we interrogated which individual PCSs drove the associations with individual phenotypes. We found striking patterns of different phenotypes being largely driven by few PCSs in the model, such that some phenotypes are associated with specific genetic patterns captured by distinct genomic ICs. Some other phenotypes are associated with multiple PCSs suggesting a widespread polygenic influence on these phenotypes, associated with genetically distinct brain features. The full results of genomic ICs (figures S5 - S36 & S69 - S78) and PCs (figures S37 - S68 & S81 - S88) are presented in the supplement and are published on a publicly available website for future reference for other researchers (genomica.info). Given the large number of phenotypes, each paired with 16 ICs and 16 PCs, we here only highlight several global observations.

While many PCSs are significantly associated with a wide range of neuroimaging features, some individual PCSs clearly drive the associations with specific white matter tracts or regions of interest. Diffusion-weighted white matter metrics are significantly predicted by all PCSs (figure 3), but different components are associated with different white matter tracts, sometimes very distinctly. For example, PCS7 is strongly associated with the *Fornix* and *superior fronto-occipital fasciculus* (figure S75), while PCS2 with the *anterior thalamic radiation* (figure S74). *Subcortical volumes* are similarly predicted with a high degree of PCS specificity, such as PCS 10 markedly driving associations in *whole brainstem volume*, *Pons*, *Midbrain* and *Medulla* (figure S72); while

PCS 7 predicts multiple specific thalamic nuclei to the exclusion of all other PCSs (figure S72). Notably, the entire category of *cortical thickness* phenotypes is mainly associated with two specific components: PCS 3 and PCS 8 (figure S70). Overall our analyses show that individual PCSs capture mutually independent genetic signals, with a degree of specificity for particular tissues, following known neuroanatomical boundaries. A full overview of results can be viewed in the supplement (figure S71 – S80).

In non-neuroimaging space, associations of PCSs with physical attributes such as *height*, *weight*, *fat-mass*, *fat-free mass*, and *metabolic rate* is pervasive across all PCSs. Widespread associations with cardiovascular health metrics also stand out. Hypertension in particular is significantly associated with 13 out of 16 PCSs Most hypertension-associated PCSs are additionally, and reliably in the same direction, associated with other cardiovascular conditions such as *Angina*, *heart attack* or *stroke* (1, 5, 6, 7, 9, 11, 15, 16), and with *heart disease* of mother, father or siblings (4, 5, 6, 7, 9, 10, 12, 15, 16). Also markedly stable across components is *high Cholesterol*, which is only significantly associated in PCSs also associated with cardiovascular health outcomes, and reliably in the same direction (1, 5, 7, 11, 12, 14, 15). Please refer to the supplement table S6 and figures S5 - S36 for full statistics and figures.

Multiple other multifactorial physical health diagnoses are significantly predicted by PCSs. These diagnoses are *Nephritis* (11, 12, 16), *Osteoporosis* (4, 9, 12), *Osteoarthritis* (1, 2, 3), *Malabsorption/Coeliac disease* (5, 6, 8, 14, 15), *Hypothyroidism/Myxoedema* (2, 5, 6, 11, 13), *Asthma* (4, 6, 13), *chronic Bronchitis* (4, 12, 13), and *Diabetes* (4, 5, 10) (figures 6 & 8).

Several PCSs link to *lifestyle and behavior*, including *dietary* metrics that reliably co-associate with cardiovascular health outcomes. For instance, PCS (3, 4, 5, 6, 7, 11, 12, 13) show inverse patterns of correlations between cardiovascular health risk and the intake of healthy fats, as *oily fish intake*. Other lifestyle metrics, including *smoking*, *alcohol intake* and *activity*, co-associate across several component PCSs with cardiovascular health outcomes in the expected direction (e.g. less smoking, less alcohol intake, more activity correlated with better cardiovascular health) (1, 6, 9, 10, 11, 13, 14). Few PCSs show unexpected associations, where associations with e.g. positive cardiovascular health outcomes co-associate with increased levels of smoking or increased alcohol intake (4, 12, 15).

PCSs further significantly associate with several socioeconomic measures such as *average household income* (1, 2, 4, 7, 8, 9, 11, 12, 13, 15), *financial satisfaction* (1, 2, 11, 12, 13, 14), *owning or renting accommodations* (8, 11, 14), and the *number of vehicles owned* (8). As expected, components associated with higher *average household income* also co-associate with higher *financial satisfaction*, *owning* rather than *renting*, and the *number of vehicles* owned. PCSs displaying an association with higher average household income consistently co-associate with an increased *overall health rating* (1, 8, 9, 11, 13).

In the category of *well-being* most of the significant associations were identified with *overall health rating* (1, 5, 6, 8, 9, 11, 13, 14, 15, 16), followed by *general happiness with own health* (1, 5, 9, 14). *Happiness* was significantly predicted by PCS 2 and 13, but did not reflect a clear pattern with other health ratings, nor a clear pattern with other metrics in this category. Other mental health outcome metrics that were significantly associated with PCSs are *seen general practitioner (GP) for nerves, anxiety, tension or depression* (2, 5, 6), and *seen psychiatrist for nerves, anxiety, tension or depression* (2, 5, 8). No significant associations were found for mental health or neurological diagnostic variables included in the analysis, such as *Dementia*, *Anxiety*, *Depression*, or *Alzheimer's Disease*.

Representative examples of selected individual genomic component polygenic scores

Here we present more detailed results for 2 components for illustration. We made our selection based on interpretability and consistency of the component PGS associations across phenotypes, which does not discount the interpretability and level of insight gained from other components. Please refer to the supplementary material for the full results of the other components (table S4 – S9), or consult our website at genomICA (genomica.info) for an interactive format.

Genomic IC-PCS 15 displays a clear pattern of association with cardiovascular features (figure 6). This PCS is positively highly significantly associated with *not being diagnosed with cardiovascular problems*, and negatively associated with *diagnosed high blood pressure*. Furthermore, PCS 15 significantly predicts *angina*, *stroke*, *heart attack, diabetes*, and *high cholesterol*, and with their occurrence in family members of participants. Positive associations that fit this general theme were found in an *increased overall health rating* predicted by this PCS. This trend is also reflected in the analyses of continuous non-neuroimaging phenotypes where the PCS is negatively associated with traits in the domains of *physical health* and *physical attributes* (figure 7).

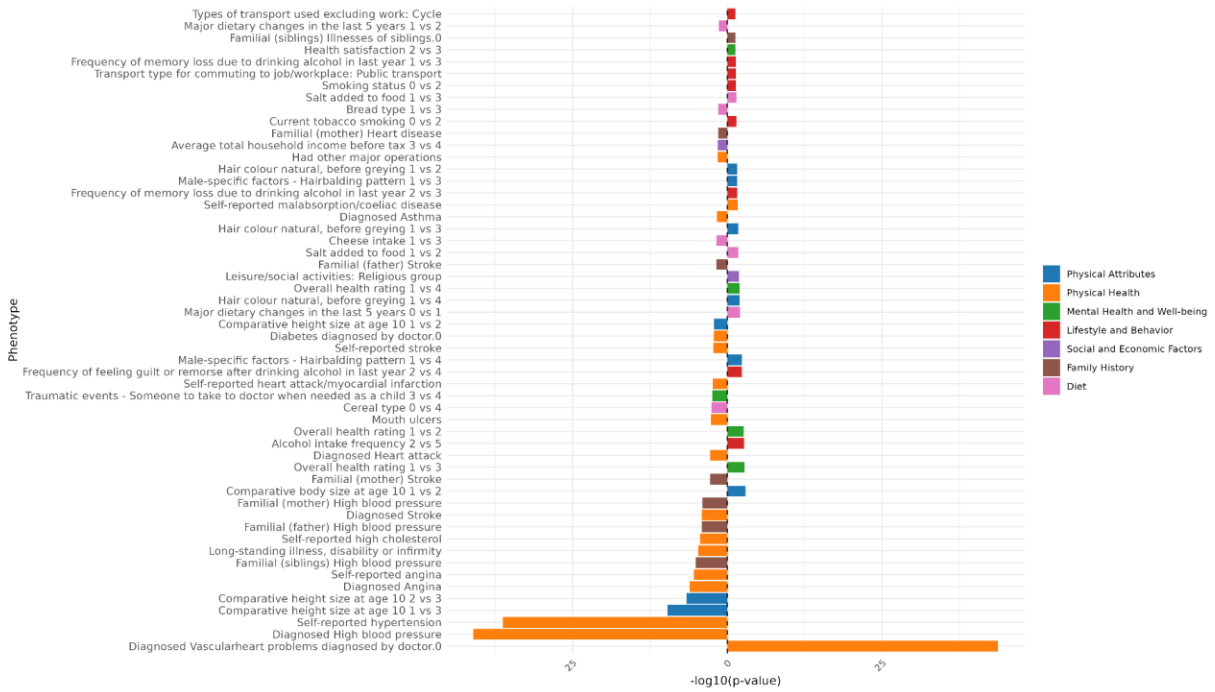

Figure 6: Overview of PCS15 significant associations with binary/binarized non-imaging phenotypes. The x-axis shows the -log(10) p-value. The direction of the bar indicates the direction of association, in this case the sign of the log-odds ratio. Right facing bars show a positive association, left facing bars a negative association. In binarized ordinal metrics the direction of the bar indicates the same, but the contrast is constructed in such a way that the lower number (e.g. 1 vs 3) is the reference. Phenotypes are binned into domains which are indicated in a color scheme contained in the legend on the right hand side of the figure.

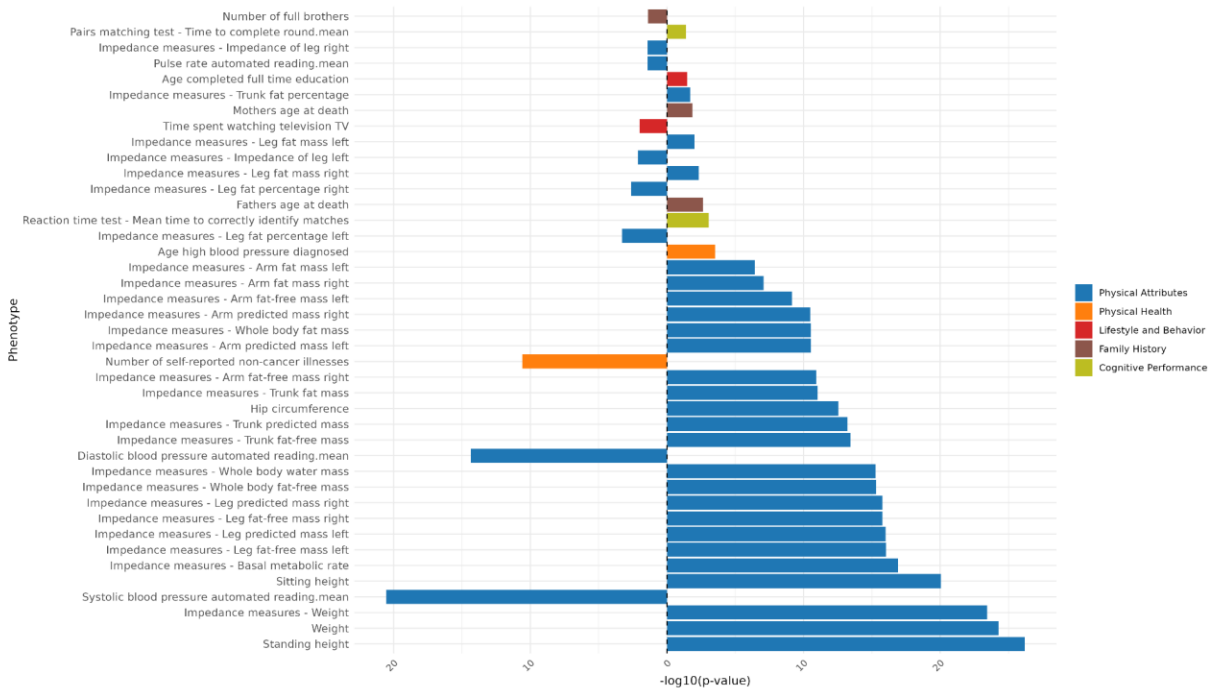

Figure 7: Overview of PCS15 significant associations with continuous non-imaging phenotypes. The x-axis shows the -log(10) p-value. The direction of the bar indicates the direction of association, in this case the sign of the beta. Right facing bars show a positive association, left facing bars a negative association. Phenotypes are binned into domains which are indicated in a color scheme contained in the legend on the right hand side of the figure.

In neuroimaging space PCS 15 drives associations with multiple dMRI derived metrics of *isotropic water fraction (ISOVF)* around the *superior* and *inferior longitudinal fasciculi* (figure 8, table S2). PCS15 uniquely captures variance in these particular neuroimaging with high significance, contrasting with other PCSs that drive associations with different dMRI phenotypes and white matter tracts (such as PCS7 driving associations in the *Fornix*).

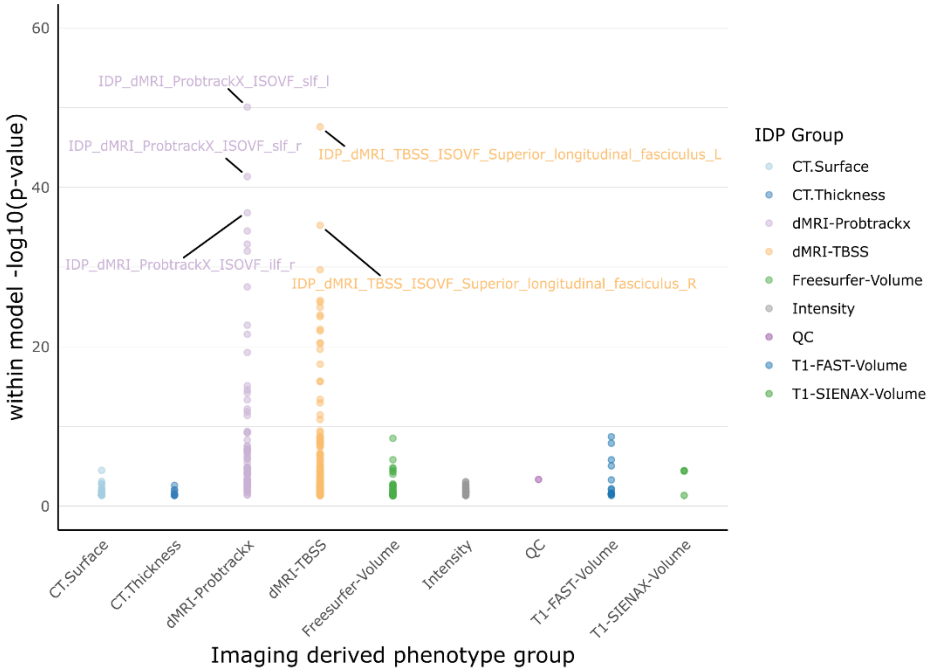

Figure 7: Dot-plot showing PCS15 associations with IDPs. The y-axis shows the -log(10) p-value, therefore the height of the dot in the plot is proportional to its significance. Each dot represents one IDP. IDPs were binned into categories shown in the legend on the right hand side of the figure, indicated by a color scheme.

Another interesting PCS to highlight is PCS 8, which is not associated with any cardiovascular health metrics in UKB participants, but is significantly associated with several *behaviour and lifestyle*, and *social and environmental* phenotypes. We find that this PCS co-associates with a wide range of dietary choices. PCS 8 shows significant associations with *eating no sugar*, tending towards *whole grain bread* and *cereal*, *less salt added to food,* and indicating a *major dietary shift within the last 5 years* (figure 9). Furthermore, we find significant associations with less *smoking* across multiple metrics, and an increased *usual walking pace*. This is combined with associations indicating an increased *average household income*, increased *financial satisfaction*, and a higher *overall health rating*. Conversely, this PCS also associates with increased levels of a*lcohol intake in both frequency* and *past alcohol consumption*, and less deliberate *physical activity in the last four weeks*. Interestingly, we also found a positive association with *maternal smoking around birth*, and a negative association with being *breastfed as a baby* (figure 9). For continuous non-imaging metrics we find the bulk of significant associations with lower *height*, *weight* and several *impedance measures*, but also an increased *impedance of the whole body* and increased *body fat percentage* (figure 10). Furthermore, there are significant associations with spending more *time watching television*, but also with spending more *time outside in both summer* and *winter*. Overall this PCS drives *dietary* and *lifestyle* associations to the exclusion of other metrics of *physical health*, indicating a pleiotropic neuroimaging-based pattern of genetic effects that act in this domain.

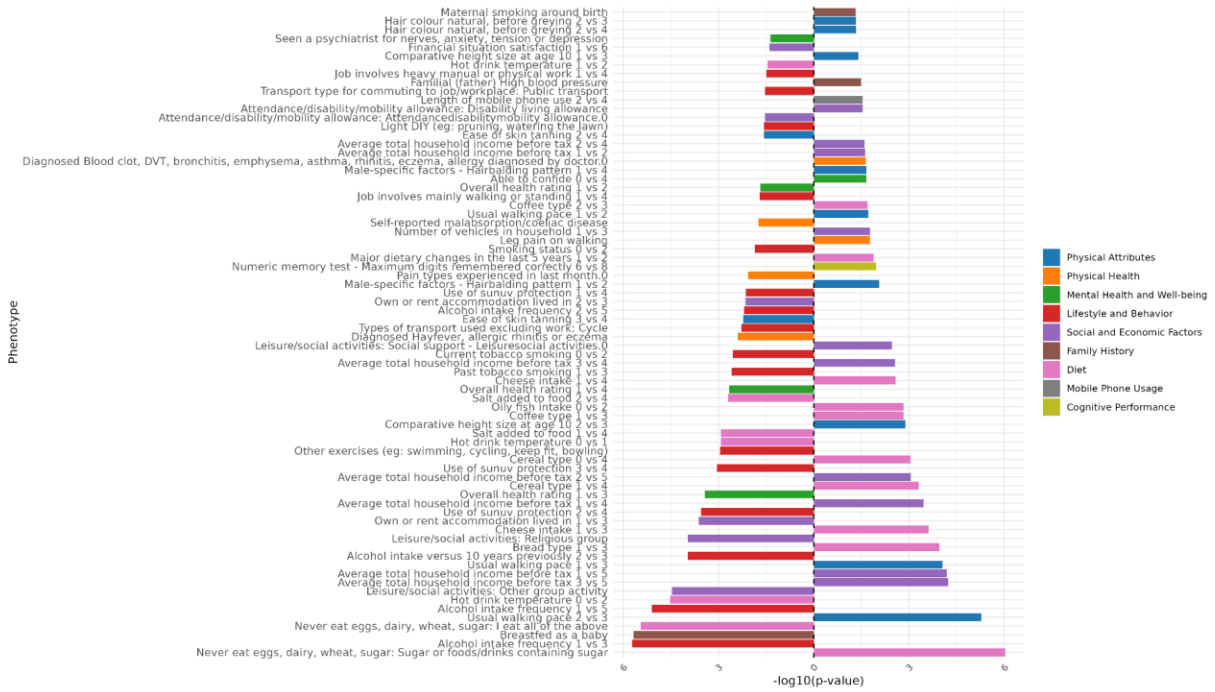

Figure 9: Overview of PCS8 significant associations with binary/binarized non-imaging phenotypes. The x-axis shows the -log(10) p-value. The direction of the bar indicates the direction of association, in this case the sign of the log-odds ratio. Right facing bars show a positive association, left facing bars a negative association. In binarized ordinal metrics the direction of the bar indicates the same, but the contrast is constructed in such a way that the lower number (e.g. 1 vs 3) is the reference. Phenotypes are binned into domains which are indicated in a color scheme contained in the legend on the right hand side of the figure.

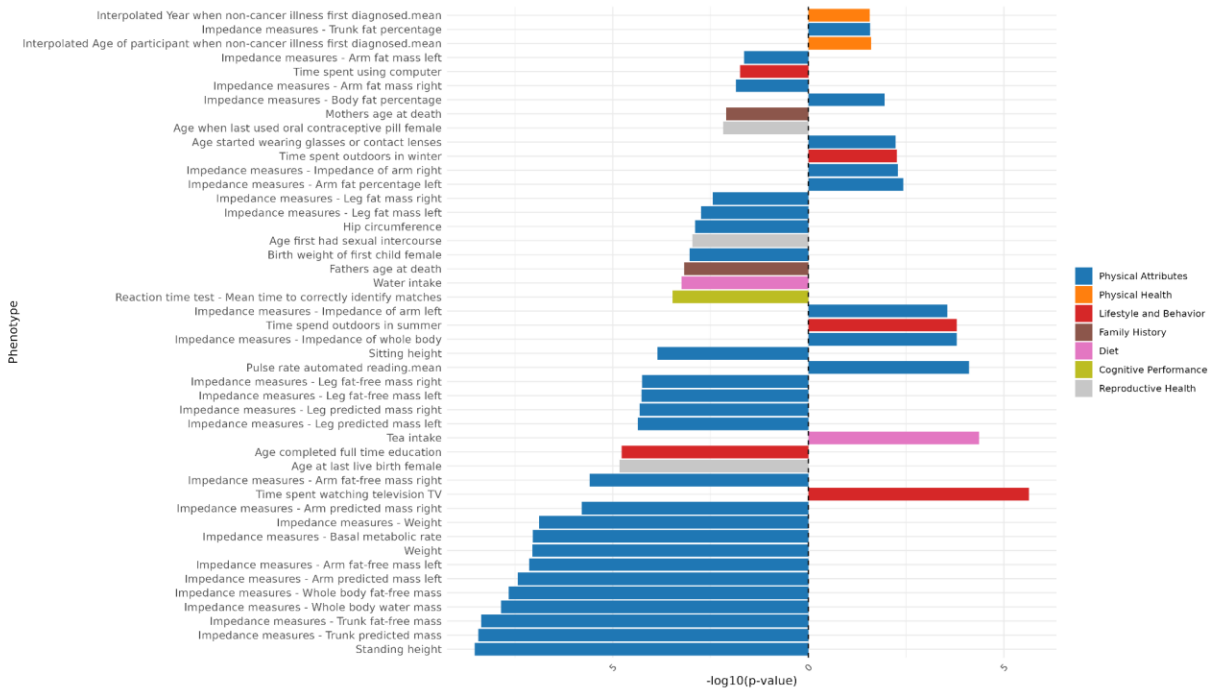

Figure 10: Overview of PCS8 significant associations with continuous non-imaging phenotypes. The x-axis shows the -log(10) p-value. The direction of the bar indicates the direction of association, in this case the sign of the beta. Right facing bars show a positive association, left facing bars a negative association. Phenotypes are binned into domains which are indicated in a color scheme contained in the legend on the right hand side of the figure.

In neuroimaging space PCS 8 captures the bulk of phenotypic variance in the domain of *cortical thickness*, and drives associations across a wide range of brain imaging phenotypes across hemispheres together with PCS 3 (figure S70). In other neuroimaging categories this PCS is notably inconspicuous, with the exception of *cortical* and *subcortical volumes*, and *intensity*, where it contributes to, but is not the main driver of them (figure 11, figures S71-S73). Taken together, this suggests that IC 8 captures polygenic effects chiefly related with *cortical thickness* that seem to link to many *dietary*, *lifestyle and behaviour*, and *environmental* variables.

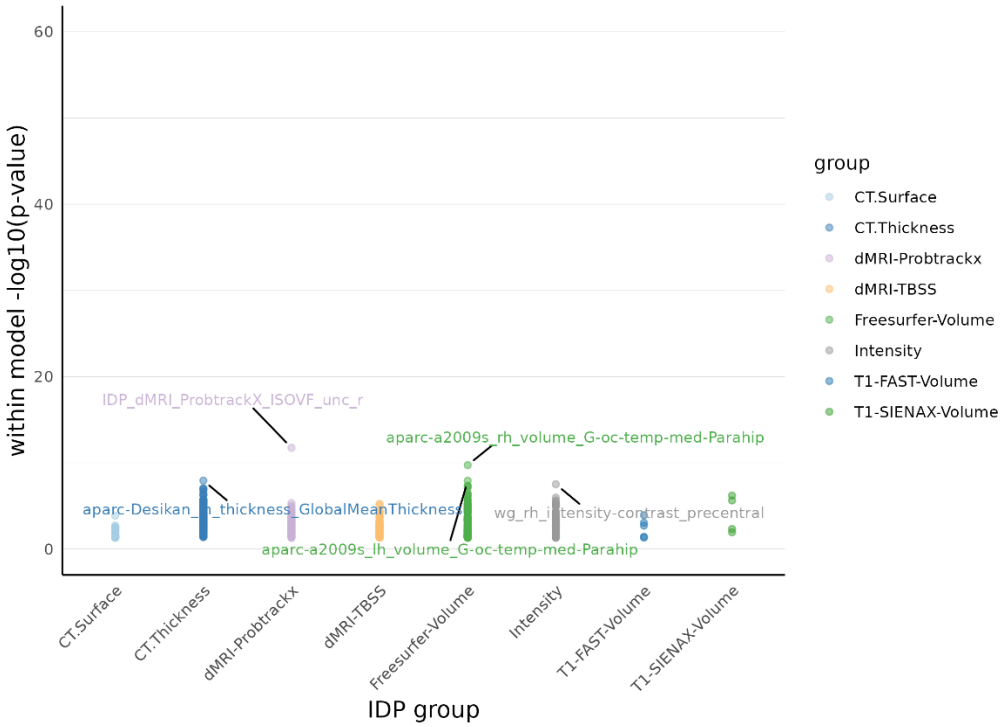

Figure 11: Dot-plot showing PCS8 associations with IDPs. The y-axis shows the -log(10) p-value, therefore the height of the dot in the plot is proportional to its significance. Each dot represents one IDP. IDPs were binned into categories shown in the legend on the right hand side of the figure, indicated by a color scheme.

Discussion

In this exploratory study we demonstrate the power and utility of genomICA derived components to congruently capture variance in an independent sample in a polygenic risk score analysis framework. We decomposed high-dimensional genome-wide brain-wide GWAS summary statistics into 16 PCs and 16 ICs that represent latent patterns of genetic effects linked to neuroimaging phenotypes across multiple categories and brain features. We calculated component-based PCSs in an independent sample of the UKB and then used them to predict a wide range of imaging and non-imaging phenotypes. These PCSs share little collinearity, confirming that polygenic scores derived from PCA and ICA decompositions reflect independent genetic effect patterns (figure 2). Our analysis also shows that PCSs predict out-of-sample neuroimaging phenotypes with substantial phenotypic variance explained (R^2_{max} = 0.12 , R^2_{mean} = 0.04). In many phenotypes this constitutes a large portion of the heritable variance (table S4 & S5). As cross-validation, we also correlated IC PCS effects (*t-values*) and initial neuroimaging-phenotype loadings in the discovery sample, which revealed that they are highly correlated and therefore reflect independent sample reliability (table S3). Further, many individual PCSs drive associations with high specificity for large white matter tracts and brain regions across several categories, thereby validating our assumption that genomICA components reflect meaningful pleiotropic patterns of genetic effects associated with distinct biological pathways. Expanding this analysis to non-neuroimaging phenotypes revealed that the neuroimaging-based PCSs significantly predict a wide range of phenotypes across many data domains albeit with much smaller effect sizes. Many PCSs display internally consistent patterns of domain specific associations, indicating joint genetic effects acting on various phenotypes within a domain, thus providing construct validity of our genomICA approach. One of these domains includes physical attributes such as height, weight, and impedance measures, which suggests an inherent link between brain-related genetic associations, and those linked to general physical attributes. Further, several PCSs show cross-domain co-associations that suggest a common, often interpretable pleiotropic genetic substrate linked to many phenotypes. All of this suggests that decomposition of GWAS summary statistics into maximally independent, pleiotropic components captures independent genetic signals, grouping individual variant effects in meaningful ways, which may have implications for the potential of these components to stratify existing cohorts. This potential stratification power may lie in identifying how distinct PCS profiles relate to variation in phenotypic outcomes across individuals.

Compelling co-association patterns between cardiovascular health outcomes and *socio-economic* and *lifestyle* metrics are dominant across PCSs. High cholesterol is a well-known risk factor for heart disease[38], and in the presented PCSs show a stable pattern with directly proportional associations of cardiovascular health outcomes. This suggests a common, heritable brain-related genetic correlate that links levels of cholesterol and the risk for heart diseases. This interpretation

is for example illustrated by the co-association with *oily fish intake*, which are rich in unsaturated fats, and has been shown to lower cholesterol levels and cardiovascular health risk[39]. Across PCSs we find stable inverse correlations of oily fish intake and high cholesterol, and additionally a wide-spread pattern of inverse correlations of oily fish intake and cardiovascular health outcomes. Cholesterol plays a critical role during white matter myelination[40] and is highly concentrated within the brain at 15% - 25% of body cholesterol[41]. Present PCSs significantly associated with high cholesterol additionally show their strongest imaging space associations in white matter features. Furthermore, white matter hyperintensities have been linked to an increased risk for stroke[42], and Hypertension is a well-known risk factor for white matter lesions[43], all of which suggest a strong link between cardiovascular health outcomes and white matter structure and integrity. This is compounded by co-association patterns that link cardiovascular health and *smoking* behaviour, which is inversely co-associated across several PCSs (1, 6, 9, 10, 11, 13, 14), and has been shown to be a risk factor for adverse cardiovascular outcomes[44]. This suggests that presented genomICA components capture large amounts of gene-by-brain variance related to cardiovascular health and brain-related risk factors for cardiovascular health, which is consistent across components and dominates PCS associations. This could be an indicator for the relative effect sizes of these risk factors as compared to more subtle gene-brain interaction patterns not picked up by our association analysis. Overall, genomICA based PCSs demonstrate a highly consistent, data-driven avenue of predicting linked risk of cardiovascular health and related risk factors.

In the domain of *social and environmental* variables PCS associations display marked internal consistency. An increased *average household income* is consistently co-associated with *financial satisfaction*, *owning rather than renting* accommodations and an increased *number of vehicles* owned, and *vice versa*. Additionally, well-being measures of *overall health rating* are consistently predicted by PCSs that also predict *average household income* in the same direction, a connection that has been widely accepted in social study literature[45–47]. This suggests that brain-based pleiotropic patterns captured across components consistently reflect an environmental link between genetic signal associated with brain phenotypes, socioeconomic metrics and *overall health rating*.

PCS8 stands out through its unique association pattern, which does not include cardiovascular features and is driven by *diet*, *lifestyle and behaviour*, and *social and environmental variables*. The high specificity towards metrics from these domains, combined with the wide range of associations in the imaging space, suggests that this component captures strong gene-environment correlations that act on brain function and structure[23].

In separate work, our group applied gene-enrichment analyses using FUMA[48] and MAGMA[49] to the same components[50]. This work indicated that genomICA components capture an array of biological processes, which includes the components 8 and 15 discussed in detail above. IC8

showed enrichment for zinc/calcium signaling (SLC39A8, DOC2A) and redox balance (MSRB3), combined with developmental regulation (PPP4C, YPEL3). This is compounded by IC8 trait associations matching genetic variants previously associated with diet, alcohol intake and smoking[50]. IC15 was found to capture genetic signal related to white matter organisation through immune and lipid metabolic pathways[50]. Particularly a dominant locus in the HLA region stands out among other loci involved in immune functioning/neuroinflammation, and myelination, which have been implicated to impact brain white matter phenotypes [51] and cardiovascular health outcomes[52] respectively. Furthermore, matching trait associations of the original IC15 include healthy fat intake, cholesterol metrics, diabetes type 2, and blood pressure metrics[50]; genomica.info). Both examples match the co-association patterns of PCS8 and PCS15 well and strengthen the interpretation that they capture relevant biological signals out-of-sample.

When comparing outcomes of PCSs derived from genomic ICs and genomic PCs, qualitative differences appear minor. Genomic PC-PCSs show strong correlations between PCS effect sizes (*t-values*) and component trait-loadings, demonstrating high intersample reliability (table S3). Both approaches explain similar magnitudes of variance across neuroimaging and non-neuroimaging domains, and each yields components driving associations within specific neuroimaging categories (table S5). In non-neuroimaging space, genomic PC-PCSs are associated with distinct patterns across *cardiovascular health*, *lifestyle and behaviour*, *diet*, and *socioeconomic factors* (figures S37 - S68). While genomic ICs capture maximally independent genetic patterns and genomic PCs capture orthogonal axes of maximum variance, both approaches reveal compelling patterns of co-association. We expect genomic ICs to be more interpretable and specific, while PC1 and PC2 reflect mixtures of dominant genetic effects, with higher-order PCs capturing less interpretable variance. Further investigation into the differences between PCs and ICs is warranted, given the richness of findings from genomic PCs alone.

While PCSs convincingly captured variance in physical health and environmental variables, associations with mental health were limited, and no significant associations with formal mental health diagnoses were identified. This suggests that focusing on neuroimaging genomics may not be the optimal strategy for uncovering the genetic mechanisms underlying mental or neurological phenotypes. Instead, the dominant genetic influences on individual variation in brain features appear to relate more strongly to physical characteristics—particularly body shape, height, and cardiovascular traits—and, to a lesser extent, to environmental and lifestyle factors. In contrast, genetic overlap with neurological or psychiatric diagnoses, as well as with broader mental health traits, appears minimal. One notable exception was a broad depression category defined as "ever visited a GP for stress or anxiety," which did show some association, though it remains an outlier.

There are limitations to be aware of. Our proposed approach is limited by GWAS summary statistic discovery sample size. The present data allows for decomposition into 16 stable,

multivariate ICs which reflect 16 maximally independent pleiotropic patterns across involved summary statistics[34]. The multitude of genetically driven mechanisms influencing brain phenotypes is unlikely to be captured in only 16 components. This is reflected in the fact that our components explain ~39% of the SNP-effect variance across summary statistics. The remaining ~61% of genetic signal remains unaccounted for in our components, as it was too noisy to capture reliably with the present data[34], but it could include a substantial amount of genetic variation related to mental health outcomes. Furthermore, it is likely that the ~39% of the SNP-effect variance reflects a larger mixture of intermediate processes than only 16, therefore each component may reflect multiple underlying processes. Another limitation is the foundation of genomICA, GWAS summary statistics themselves, which are observational data that inherently limit the causal inferences one can make from them and their downstream analyses. With these limitations in mind, the interpretation of the presented multivariate PCS analysis is tentative. Nevertheless, the presented PCS associations show clear patterns of co-association within and across components that are supported by existing literature and gene-enrichment analysis of the components[50]. The present exploratory analysis of genomICA derived components thus provides tentative evidence for common, brain-related genetic pathways linked to various risk factors for cardiovascular health, lifestyle and even environmental influences. Several phenotypes and components have not been discussed in the main text, and additional investigations into the nature of these components are warranted. Overall, the outcome and interpretation of this analysis is hypothesis generating and should be thoroughly tested in further downstream analyses to confirm presented findings.

By expanding the decomposition to include environmental, sociodemographic, lifestyle, and mental health related phenotype GWASs in the future, we hope to improve the sensitivity of PGS analysis to capture variance that is driven by covariance between brain phenotypes, environment and mental health outcomes simultaneously.

In conclusion, our novel GWAS summary statistic decomposition genomICA has yielded data-driven, uncorrelated polygenic scores that reliably predict individual differences in a wide range of physical and behavioural phenotypes out-of-sample. The present analysis revealed that clear patterns of associations emerge that are supported by literature, particularly between cardiovascular health outcomes and lifestyle, and socioeconomic factors and overall health ratings. Overall, the presented results strengthen the idea of gene-environment correlations that jointly act on brain structure and function. These results are promising for future applications of genomICA and subsequent polygenic component scores. Genomic components are uncorrelated and unmix intermediate genetic processes. Therefore they may help stratify the (etiological) heterogeneity of both clinical and general population cohorts. In future work, we will investigate how individual PCS profiles differ and how they may predict outcomes to assess their stratification performance. With sufficient sample sizes, genomICA is applicable to domains of

environmental and other -omics data, which may further increase its potential to reconstruct independent genomic signal from noisy data and thereby elucidate biological pathways subserving complex traits.

Methods

Methods Overview Flowchart

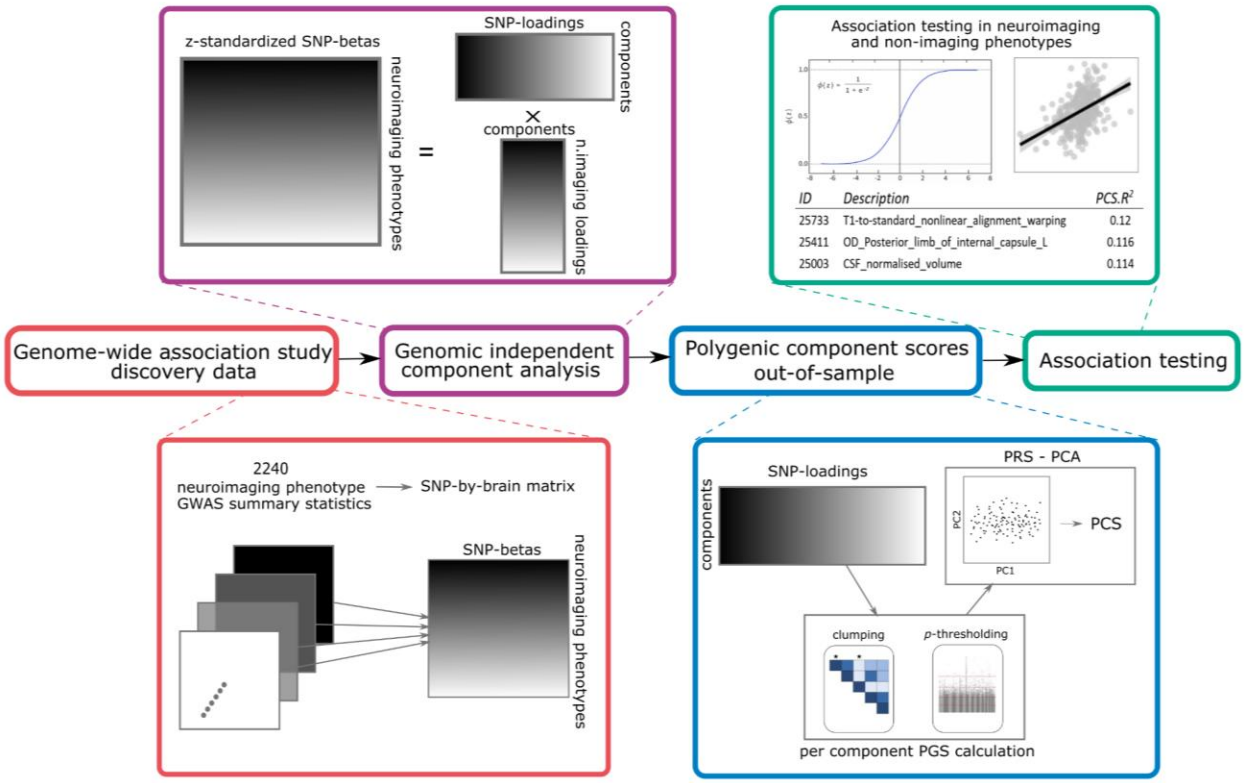

Figure 1: Schematic overview of the genomic independent component analysis with subsequent polygenic scoring analysis.

Data

The analysis required multiple data sets. First, a discovery dataset of SNP-by-brain phenotype GWAS summary statistics for the calculation of latent genomic components. Second, two out-of-sample target datasets for calculation of PCSs: one with neuroimaging data (N = 3,506), and one without neuroimaging data (N = 338,429). As the discovery dataset (N = 33,224), we used the Oxford Brain Imaging Genetics (BIG-40) database containing the results of over 4000 GWASs of MRI based derived brain phenotypes in the UK Biobank (UKB)[8,53]. The neuroimaging traits were grouped into categories: T1-weighted MRI, diffusion-weighted MRI (dMRI), susceptibility-weighted imaging (SWI), fluid-attenuated inversion recovery (FLAIR), task-based functional MRI (tfMRI), resting-state functional MRI (rsfMRI), as well as a quality control (QC) category. Metrics of rsfMRI with relatively high heritability were the amplitudes of functional signal fluctuations and six ICA-derived 'global' measures of functional connectivity and are therefore included in the analysis[1,8]. Excluded brain phenotypes with low heritability were node-to-node rsfMRI metrics[1]. Specifics on the UKBB GWAS pipeline can be found in the main publications[1,8] and the BIG-40

website (https://open.win.ox.ac.uk/ukbiobank/big40/). This set of N = 2,240 neuroimaging GWAS summary statistics were used to extract latent genomic components using genomICA, described in a subsection below.

To compute PCSs based on genomICA derived principal and independent components we used an out-of-sample set of UKB participants (i.e. not included in genomICA decomposition). We were granted access to the full UKBB genotype data of N = 377,663 participants aged between 40 and 69 years. These data were genotyped using the Affymetrix UK BiLEVE Axiom array or the UK Biobank Axiom array. Of the 377,663 participants originally available, 33,224 were excluded due to their inclusion in the discovery sample used to derive the genomICA components. This left a final genotype dataset of 344,439 participants and 10,203,392 SNPs (post-QC) available for polygenic scoring prior to clumping. Among these, 3,506 participants had neuroimaging data and were added to the UK Biobank after 2021, ensuring no overlap with the discovery sample. These 3,506 individuals constituted the sample for the first analysis, and were removed from the dataset for the second analysis. For the second analysis of non-neuroimaging phenotypes the paper of Watanabe and colleagues (2019) was consulted to select non-neuroimaging phenotypes with clean data and large sample size available in UKB. Based on the data availability in our UKB data the second independent target dataset for non-neuroimaging traits consisted of N = 338,429 participants across non-imaging phenotypes.

Data quality control of target samples

The initial target sample genetic dataset contained 93,095,623 SNPs[54]. The data was quality controlled for sex mismatches, sample missingness > 0.05, heterozygosity outside of 3 standard deviations (SD), and participants of non-European descent (self-report) were removed. Principal components (PC) to account for population stratification during association testing were derived. Further, relatedness was accounted for by calculating the KING kinship coefficients and creating family clusters for participants with KING > 0.0884 (first – and second-degree relatives). Then, the sample was reduced until only one participant was left per family cluster[55]. SNPs were removed if they had Hardy-Weinberg equilibrium (HWE) < 1e-6, genotype missing rate > 0.05, minor allele frequency (MAF) < 0.005, or imputation quality INFO score < 0.8. After quality control the final SNP count was N = 10,203,392[55].

The neuroimaging target dataset of 3,506 participants was quality controlled by checking if the field values were consistent with the neuroimaging-phenotype they represent via the UKB catalogue. Then, the number of missing values per phenotype was assessed to ensure that phenotypes had sufficient datapoints for regression analysis (N > 2,500). After QC, neuroimaging-phenotypes (N = 2,240) were filtered based on their loadings in the neuroimaging-space of independent and principal genomic components (|z| > 2). This brought the number of

neuroimaging-phenotypes down to N = 1,269 for genomic ICs and N = 1,330 for genomic PCs (tables S4 & S5).

The non-imaging target dataset consisted of 152 continuous and 758 binary, categorical or ordinal phenotypes from the UKB data repository[56]. The data consisted of phenotypes from multiple domains: *physical attributes*, *physical health*, *reproductive health*, *mental health and well-being*, *lifestyle and behaviour*, *social and economic factors*, *cognitive performance*, *diet*, *family history*, and *mobile phone usage*. A full list of phenotypes used in the analysis is shown in supplement (table S6 & S8). We altered the data coding for some variables based on the recommendations of Watanabe and colleagues ([56] - supplementary table 1). In the binary, categorical and ordinal phenotypes we also removed all instances where participants answered "do not know", "do not recall", "prefer not to answer", or "less than one [item] a week" across all phenotypes. Lastly, we only retained those phenotypes that had >50,000 data points, as in[56], to ensure sufficient power. Ordinal phenotypes were binarized in a pair-wise manner so that the full range of answer contrasts could be analysed. Categorical multiple choice answers were binarized. For continuous phenotypes with multiple measurements at the same visit (e.g. systolic blood pressure) the mean across measurements was calculated.

Clumping

Prior to the decomposition with genomICA the GWAS summary statistic SNPs (N = 17,103,079) of 2240 neuroimaging phenotypes were clumped at LD-threshold r^2 = 0.5 and significance threshold p = 0.001 with a 250kb range using plink2.0[57]. This resulted in N = 1,032,967 lead-SNPs in the genome-dimension included in the decomposition. The data was clumped in such a way that SNPs in the European 1000G reference panel are prioritized and retained to facilitate overlap with other GWAS summary statistics where the 1000G reference panel is often used. During post-hoc QC, 440 SNPs were identified as duplicates and therefore removed, bringing the final number of SNPs in the SNP x brain data to N = 1,032,527.

GenomICA components

GenomICA was applied as previously described in [34]. The GWAS summary statistics were concatenated into a *m x n* matrix, representing a genome-wide matrix of MRI-derived neuroimaging phenotypes (m = 2240) and common genetic variants (n = 1,032,527 clumped SNPs). Multivariate decomposition was applied to the z-standardized GWAS betas, such that SNPs are normalized for their own standard errors (SE). The matrix was decomposed using v3.15 of Multivariate Exploratory Linear Optimised decomposition into independent components (MELODIC), a probabilistic ICA algorithm[58]. We applied the mask of clumped SNPs to the *m* x *n* matrix such that only lead-SNPs remain. Then, MELODIC was applied to the matrix, decomposing it into a predefined number of PCs. The PCs are then rotated to optimize non-Gaussianity across

components, thereby creating the same number of independent components (IC). Each component is represented by two vectors: one vector of SNP-loadings and a vector of neuroimaging-phenotype-loadings. This feature of genomICA decomposition allows the linking between patterns of genetic effect sizes and corresponding patterns of brain features. In our previous work we determined that decomposing into 10 components was optimal for reproducibility for N = ~160,000 SNPs clumped at $r^2 < 0.1$. For this highly clumped dataset 10 components captured ~39% of SNP effect variance. Here, with less stringent clumping 16 PCs and ICs (Dimension 16) captured the same (reproducible) ~39% SNP effect variance. For this reason, we here chose to decompose into 16 components, to capture the same, previously determined reproducible, variance.

The final output of genomICA is therefore 16 PCs and 16 ICs, each of which is a vector of SNP-loadings and trait-loadings. Both PCs and ICs are used in subsequent analyses.

Polygenic Score calculation

To compute polygenic component scores (PCSs), we applied PRS-PCA[59], using SNP-loadings from genomic components as weights. As previously described, polygenic scores were first calculated using PRSice 2.3.5[25] across a range of p-value thresholds (0.001, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, and 1). SNPs were re-clumped using an LD threshold of $r^2 < 0.1$ within a 250 kb window, and a sliding window approach, reducing the set to N = 290,660 lead SNPs. Then, PCA was applied to identify the principal axis of variance across thresholds. The first principal component effectively captured the shared variance across thresholds for all components (table S1 & S2). This component was used for all downstream analyses, and is henceforth referred to as the PCS of a given component. We opted not to use methods like PRS-CS or LDpred[60,61], which rely on continuous shrinkage and LD-informed priors, due to insufficient LD structure in our genomic components.

Polygenic Component Score association testing

PCSs of all ICs and PCs were used as predictors in general linear regression models (for continuous phenotypes) and logistic regression models (for binary/binarized phenotypes). To calculate the phenotypic variance explained by PCSs we constructed a full-model with all 16 PCSs and covariates, and a null-model, with only covariates as predictors. For all models we calculate the *conventional* and *proportional* variance explained (R^2). Although both are routinely reported in polygenic scoring they can differ substantially depending on the variance explained by the covariates, so for comparison with other literature, we here report both. Conventional ($R^2_{conventional}$) is the variance explained by the null model subtracted from the variance explained by the full model.

$R^2_{conventional} = R^2_{full} - R^2_{null}$

Proportional variance explained is the proportion of the variance that the full model explains which is not captured by the null model.

$R^2_{proportional} = (R^2_{full} - R^2_{null}) / (1 - R^2_{null})$

For logistic regression models we calculated *Nagelkerke's* pseudo-R^2, as well as *McFadden's* pseudo-R^2 as a comparative metric, and then calculated the proportion of phenotypic PCS-R^2 in the same way as outlined above. To determine the significance of PCS-R^2 we ran an ANOVA between the full and null general linear models, and a χ^2-test for logistic binomial models. *P*-values were corrected for type I error with false-discovery rate (FDR) correction across outcome phenotypes tested. Finally, coefficients and p-values of individual PCSs were extracted from each model in turn to interrogate individual component associations. P-values within each model were likewise corrected for type I errors by applying FDR correction.

Covariates included in our analysis were age, sex, and body-mass-index, assessment location, genotyping batch and 40 population stratification PCs derived during data quality control. The latter provide an estimate for global genetic structure and are therefore useful for inclusion in statistical models. For the neuroimaging phenotype models we also included head-motion and head-location (x-, y-, z- axes) in the MRI-scanner as covariates.

References

1. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
2. Eyler, L. T. *et al.* Genetic and Environmental Contributions to Regional Cortical Surface Area in Humans: A Magnetic Resonance Imaging Twin Study. *Cereb. Cortex N. Y. NY* **21**, 2313–2321 (2011).
3. Glahn, D. C. *et al.* Genetic control over the resting brain. *Proc. Natl. Acad. Sci.* **107**, 1223–1228 (2010).
4. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science* **367**, eaay6690 (2020).
5. Jahanshad, N. *et al.* Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA–DTI working group. *NeuroImage* **81**, 455–469 (2013).
6. McKay, D. R. *et al.* Influence of age, sex and genetic factors on the human brain. *Brain Imaging Behav.* **8**, 143–152 (2014).
7. Pizzagalli, F. *et al.* The reliability and heritability of cortical folds and their genetic correlations across hemispheres. *Commun. Biol.* **3**, 1–12 (2020).
8. Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat. Neurosci.* **24**, 737–745 (2021).
9. Sprooten, E. *et al.* Common genetic variants and gene expression associated with white matter microstructure in the human brain. *NeuroImage* **97**, 252–261 (2014).
10. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex

traits. *Nat. Genet.* **51**, 1339–1348 (2019).

11. Winkler, A. M. *et al.* Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *NeuroImage* **53**, 1135–1146 (2010).
12. Hibar, D. P. *et al.* Novel genetic loci associated with hippocampal volume. *Nat. Commun.* **8**, 13624 (2017).
13. Matoba, N., Love, M. I. & Stein, J. L. Evaluating brain structure traits as endophenotypes using polygenicity and discoverability. *Hum. Brain Mapp.* **43**, 329–340 (2022).
14. Tissink, E. P. *et al.* Abundant pleiotropy across neuroimaging modalities identified through a multivariate genome-wide association study. *Nat. Commun.* **15**, 2655 (2024).
15. Wendt, F. R., Pathak, G. A., Tylee, D. S., Goswami, A. & Polimanti, R. Heterogeneity and Polygenicity in Psychiatric Disorders: A Genome-Wide Perspective. *Chronic Stress* **4**, 2470547020924844 (2020).
16. Allegrini, A. G. *et al.* Genomic prediction of cognitive traits in childhood and adolescence. *Mol. Psychiatry* **24**, 819–827 (2019).
17. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
18. Karlsson Linnér, R. *et al.* Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nat. Neurosci.* **24**, 1367–1376 (2021).
19. Lee, J. J., McGue, M., Iacono, W. G. & Chow, C. C. The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genet. Epidemiol.* **42**, 783–795 (2018).

20. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
21. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).
22. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
23. Sprooten, E., Franke, B. & Greven, C. U. The P-factor and its genomic and neural equivalents: an integrated perspective. *Mol. Psychiatry* **27**, 38–48 (2022).
24. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
25. Choi, S. W. & O'Reilly, P. F. PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience* **8**, giz082 (2019).
26. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013).
27. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
28. Wray, N. R. *et al.* Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry* **55**, 1068–1087 (2014).
29. Rodrigue, A. L. *et al.* Specificity of Psychiatric Polygenic Risk Scores and Their Effects on Associated Risk Phenotypes. *Biol. Psychiatry Glob. Open Sci.* **3**, 519–529 (2023).
30. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).

31. Fürtjes, A. E. *et al.* General dimensions of human brain morphometry inferred from genome-wide association data. *Hum. Brain Mapp.* **44**, 3311–3323 (2023).
32. van der Meer, D. *et al.* Understanding the genetic determinants of the brain with MOSTest. *Nat. Commun.* **11**, 3512 (2020).
33. Lazarev, D., Chau, G., Bloemendal, A., Churchhouse, C. & Neale, B. M. GUIDE deconstructs genetic architectures using association studies. 2024.05.03.592285 Preprint at https://doi.org/10.1101/2024.05.03.592285 (2024).
34. Oblong, L. M. *et al.* Principal and independent genomic components of brain structure and function. *Genes Brain Behav.* **23**, e12876 (2024).
35. Comon, P. Independent component analysis, A new concept? *Signal Process.* **36**, 287–314 (1994).
36. Liu, S., Smit, D. J. A., Abdellaoui, A., van Wingen, G. A. & Verweij, K. J. H. Brain Structure and Function Show Distinct Relations With Genetic Predispositions to Mental Health and Cognition. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **8**, 300–310 (2023).
37. Pan, X. *et al.* Exploring the genetic correlation between obesity-related traits and regional brain volumes: Evidence from UK Biobank cohort. *NeuroImage Clin.* **33**, 102870 (2022).
38. Peters, S. A. E., Singhateh, Y., Mackay, D., Huxley, R. R. & Woodward, M. Total cholesterol as a risk factor for coronary heart disease and stroke in women compared with men: A systematic review and meta-analysis. *Atherosclerosis* **248**, 123–131 (2016).
39. Yanai, H. *et al.* An Improvement of Cardiovascular Risk Factors by Omega-3 Polyunsaturated Fatty Acids. *J. Clin. Med. Res.* **10**, 281–289 (2018).
40. Saher, G. *et al.* High cholesterol level is essential for myelin membrane growth. *Nat.*

Neurosci. **8**, 468–475 (2005).

41. Dietschy, J. M. Central nervous system: cholesterol turnover, brain development and neurodegeneration. *Biol. Chem.* **390**, 287–293 (2009).
42. Debette, S. & Markus, H. S. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* **341**, c3666 (2010).
43. Verhaaren, B. F. J. *et al.* High Blood Pressure and Cerebral White Matter Lesion Progression in the General Population. *Hypertension* **61**, 1354–1359 (2013).
44. Gallucci, G., Tartarone, A., Lerose, R., Lalinga, A. V. & Capobianco, A. M. Cardiovascular risk of smoking and benefits of smoking cessation. *J. Thorac. Dis.* **12**, 3866–3876 (2020).
45. Bonner, W. I. A. *et al.* Determinants of self-perceived health for Canadians aged 40 and older and policy implications. *Int. J. Equity Health* **16**, 94 (2017).
46. Cialani, C. & Mortazavi, R. The effect of objective income and perceived economic resources on self-rated health. *Int. J. Equity Health* **19**, 196 (2020).
47. Wang, N. *et al.* Perceived Health as Related to Income, Socio-economic Status, Lifestyle, and Social Support Factors in a Middle-aged Japanese. *J. Epidemiol.* **15**, 155–162 (2005).
48. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
49. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
50. Trevisan, N. *et al.* The Genetic Architecture of Brain Structure and Function: A Data-Driven Interpretation Using genomICA. 2025.04.07.25324950 Preprint at

https://doi.org/10.1101/2025.04.07.25324950 (2025).

51. Morita, J. *et al.* Structure and biological function of ENPP6, a choline-specific glycerophosphodiester-phosphodiesterase. *Sci. Rep.* **6**, 20995 (2016).
52. Dal Lin, C., Tona, F. & Osto, E. The crosstalk between the cardiovascular and the immune system. *Vasc. Biol.* **1**, H83–H88 (2019).
53. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166**, 400–424 (2018).
54. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
55. Shi, Y., Franke, B., Mota, N. R. & Sprooten, E. Genetic liability to major psychiatric disorders contributes to multi-faceted quality of life outcomes in children and adults. 2023.01.17.23284645 Preprint at https://doi.org/10.1101/2023.01.17.23284645 (2023).
56. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
57. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
58. Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* **23**, 137–152 (2004).
59. Coombes, B. J., Ploner, A., Bergen, S. E. & Biernacka, J. M. A principal component approach to improve association testing with polygenic risk scores. *Genet. Epidemiol.* **44**, 676–686 (2020).
60. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian

regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).

61. Mak, T. S. H. *et al.* Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).