Discovering the genetic architecture of the mind

Karlsson Linnér, R.

2019

2

Genome-wide association analyses of risk tolerance and risky behaviors in over one million individuals identify hundreds of loci and shared genetic influences

“Never was anything great achieved without danger.”

Niccolo Machiavelli

“So we shall let the reader answer this question for himself: who is the happier man, he who has braved the storm of life and lived or he who has stayed securely on shore and merely existed?”

Hunter S. Thompson

“The only way to find true happiness is to risk being completely cut open.”

Chuck Palahniuk

Based on Karlsson Linnér et al. (2019). *Nature Genetics*.

Abstract

Humans vary substantially in their willingness to take risks. In a combined sample of over one million individuals, we conducted genome-wide association studies (GWAS) of general risk tolerance, adventurousness, and risky behaviors in the driving, drinking, smoking, and sexual domains. Across all GWAS we identified hundreds of associated loci, including 99 loci associated with general risk tolerance. We report evidence of substantial shared genetic influences across risk tolerance and the risky behaviors: 46 of the 99 general risk tolerance loci contain a lead SNP for at least one of our other GWAS, and general risk tolerance is genetically correlated ($|\hat{r}_g|$ ~ 0.25 to 0.50) with a range of risky behaviors. Bioinformatics analyses imply that genes near general-risk-tolerance-associated SNPs are highly expressed in brain tissues and point to a role for glutamatergic and GABAergic neurotransmission. We found no evidence of enrichment for genes previously hypothesized to relate to risk tolerance.

Introduction

Choices in important domains of life, including health, fertility, finance, employment, and social relationships, rarely have consequences that can be anticipated perfectly. The degree of variability in possible outcomes is called risk. Risk tolerance—defined as the willingness to take risks, typically to obtain some reward—varies substantially across humans and has been actively studied in the behavioral and social sciences. An individual's risk tolerance may vary across domains, but survey-based measures of *general* risk tolerance (e.g., "Would you describe yourself as someone who takes risks?") have been found to be good all-around predictors of risky behaviors such as portfolio allocation, occupational choice, smoking, drinking alcohol, and starting one's own business[1–3].

Twin studies have established that various measures of risk tolerance are moderately heritable ($h^2 \sim 30\%$, although estimates in the literature vary[3–5]). Discovery of specific genetic variants associated with general risk tolerance could provide insights into underlying biological pathways; advance our understanding of how genetic influences are amplified and dampened by environmental factors; enable the construction of polygenic scores (indexes of many genetic variants) that can be used as overall measures of genetic influences on individuals; and help distinguish genetic variation associated with general versus domain-specific risk tolerance.

Although risk tolerance has been one of the most studied phenotypes in social science genetics, most claims of positive findings have been based on small-sample candidate gene studies (**Supplementary Table 1**), whose limitations are now appreciated[6]. To date, only two loci associated with risk tolerance have been identified in genome-wide association studies (GWAS)[7,8].

Here, we report results from large-scale GWAS of self-reported general risk tolerance (our primary phenotype) and six supplementary phenotypes: "adventurousness" (defined as the self-reported tendency to be adventurous vs. cautious); four risky behaviors: "automobile speeding propensity" (the tendency to drive faster than the speed limit), "drinks per week" (the average number of alcoholic drinks consumed per week), "ever smoker" (whether one has ever been a smoker), and "number of sexual partners" (the lifetime number of sexual partners); and the first principal component (PC) of these four risky behaviors, which we interpret as capturing the general tendency to take risks across domains. All seven phenotypes are coded such that higher phenotype values are associated with higher risk tolerance or risk taking. **Table 2.1** lists, for each GWAS, the datasets we analyzed and the GWAS sample sizes.

Results

Association analyses

All seven GWAS were performed in European-ancestry subjects; included controls for the top 10 (or more) principal components of the genetic relatedness matrix and for sex and birth year (**Supplementary Table 2**); and followed procedures described in a pre-specified analysis plan and in the **Supplementary Note**.

In the discovery phase of our GWAS of general risk tolerance (*n* = 939,908), we conducted a GWAS using the UK Biobank (UKB, *n* = 431,126) and then performed a sample-size-weighted meta-analysis of those results with GWAS results from a sample of research participants from 23andMe (*n* = 508,782). The UKB measure of general risk tolerance is based on the question:

"Would you describe yourself as someone who takes risks? Yes / No." The 23andMe measure is based on a question about overall comfort taking risks, with five response options ranging from "very comfortable" to "very uncomfortable." The genetic correlation[9] between the UKB and 23andMe cohorts ($\hat{r}_g = 0.77$, $SE = 0.02$) is smaller than one but high enough to justify our approach of pooling the two cohorts (see Section 2 in the Supplementary Note of ref. [10] for a theoretical demonstration of the merits of pooling cohorts despite moderate heterogeneity of phenotype measures).

The Q-Q plot (**Supplementary Fig. 1a**) from the discovery GWAS exhibits substantial inflation ($\lambda_{GC} = 1.41$). According to the estimated intercept from a linkage disequilibrium (LD) Score regression[11], only a small share of this inflation (~5%) in test statistics is due to confounding biases such as cryptic relatedness and population stratification. To account for these biases, we inflated GWAS standard errors by the square root of the LD Score regression intercept[12].

We identified 124 approximately independent SNPs (pairwise $r^2 < 0.1$) that attained genome-wide significance ($P < 5\times10^{-8}$). These 124 "lead SNPs" are listed in **Supplementary Table 3** and shown in **Figure 2.1a**. All have coefficients of determination (R^2's) below 0.02%, and the SNP with the largest per-allele effect is estimated to increase general risk tolerance by ~0.026 standard deviations in our discovery sample (**Supplementary Fig. 2**). To test if the lead SNPs' effect sizes are heterogeneous across the 23andMe and UKB cohorts, we generated an omnibus test statistic by summing Cochran's Q statistics across all lead SNPs; consistent with our genetic correlation estimate of less than unity between the two cohorts, we rejected the null hypothesis of homogeneity ($P = 4.32\times10^{-5}$; **Supplementary Note**). To define genomic loci around the lead SNPs, we took the physical regions containing all SNPs in LD (pairwise $r^2 > 0.6$) with the lead SNPs and merged loci within 250 kb of each other; the 124 lead SNPs are located in 99 such loci (**Supplementary Table 3**). We supplemented those analyses with a conditional and joint multiple-SNP (COJO) analysis[13], which identified 91 genome-wide significant "conditional associations" (**Supplementary Table 3**).

In the replication phase of our GWAS of general risk tolerance (combined $n = 35{,}445$), we meta-analyzed summary statistics from ten smaller cohorts. Additional details on cohort-level phenotype measures are provided in **Supplementary Table 4**. The cohorts' survey questions differ in terms of their exact wording and number of response categories, but all questions ask subjects about their overall or general attitudes toward risk. The genetic correlation[9] between the discovery and replication GWAS is 0.83 ($SE = 0.13$). 123 of the 124 lead SNPs were available or well proxied by an available SNP in the replication GWAS results. Out of these 123 SNPs, 94 have a concordant sign ($P = 1.7\times10^{-9}$) and 23 are significant at the 5% level in one-tailed t tests ($P = 4.5\times10^{-8}$) (**Supplementary Fig. 3**). This empirical replication record closely matches theoretical projections that take into account sampling variation and the winner's curse (**Supplementary Note**).

In the UKB we tested and confirmed that a much higher fraction of males (34%) than females (19%) described themselves as risk tolerant on the general risk tolerance measure (t-test $P < 1\times10^{-100}$; **Supplementary Fig. 4**), consistent with much prior research[14,15]. We used bivariate LD Score regression[12] to calculate the genetic correlation between GWAS performed separately in the sample of females and in the sample of males in the UKB. Our estimate ($\hat{r}_g = 0.822$, $SE = 0.033$) is high enough to justify our approach of pooling males and females in our other analyses to maximize statistical power[10]. Nonetheless, our estimate is significantly smaller than unity, suggesting that the autosomal genetic factors contributing to general risk tolerance, while largely similar across sexes, are not identical.

Our six supplementary GWAS—of adventurousness, the four risky behaviors, and their principal component (*n* = 315,894 to 557,923; **Supplementary Tables 4-5**)—were conducted using methods comparable to those in the primary GWAS, except that they had no replication phases and most involved a single large cohort. **Supplementary Fig. 1** shows Q-Q plots and **Supplementary Fig. 5** shows Manhattan plots.

Table 2.1 provides a summary overview of the seven GWAS. We identified a total of 864 "lead associations": the sum total of the 124 general-risk-tolerance lead SNPs together with the 740 lead SNPs from the six supplementary GWAS. (These 864 lead associations were obtained by considering each of our seven phenotypes separately and using the standard genome-wide significance *P* value threshold of 5×10^{-8}. If we instead consider the seven GWAS jointly and use a Bonferroni-corrected *P* value threshold of 7.1×10^{-9} (= $5\times10^{-8}/7$), we obtain 566 lead associations across the seven GWAS.) Since we did not have the data to conduct replication analyses of the lead associations from the supplementary GWAS, we calculated the "maxFDR"[16], a theoretical upper bound on the false discovery rate (FDR), for each GWAS. The maxFDR estimates were low across all GWAS (the highest estimate was 1.22×10^{-3}, for automobile speeding propensity), thus providing reassurance about the robustness of the lead associations.

Applying our locus definition, we identified a total of 703 "locus associations": the sum total of the 99 general-risk-tolerance loci together with the 604 loci from the supplementary GWAS (**Supplementary Note**). Pooling the loci corresponding to the 703 locus associations, and merging loci within 250 kb from each other, yields 444 distinct loci. COJO analyses[13] identified a sum total of 655 conditional associations across all seven GWAS. (If we instead consider the seven GWAS jointly and use a Bonferroni-corrected *P* value threshold of 7.1×10^{-9} (= $5\times10^{-8}/7$), we obtain 464 locus associations and 505 conditional associations across the seven GWAS.) We verified that the results of the COJO analyses are consistent with those from multiple regressions using individual-level genotype-dosage data from the UKB (**Supplementary Note**). **Supplementary Tables 3** and **6-7** report the lead SNPs, the genomic loci, and the results of the COJO analyses. **Table 2.1** also shows the SNP heritabilities[17] of the seven phenotypes, calculated from the GWAS results; the SNP heritabilities range from ~0.05 (for general risk tolerance) to ~0.16 (for the first PC of the four risky behaviors).

We note that 212 of the 864 lead associations are located within long-range LD regions[18] or candidate inversions (i.e., genomic regions that are highly prone to inversion polymorphisms; **Supplementary Note**). Of these, only 109 are also conditional associations, and 46 are in loci that contain no conditional associations, thus indicating that many lead associations in the long-range LD regions or candidate inversions may tag causal variants that are also tagged by other lead associations. We discuss some of these regions in the next section.

Genetic overlap

There is substantial overlap across the results of our GWAS. For example, 46 of the 99 general-risk-tolerance loci contain a lead SNP of at least one of the other GWAS, and 72 of the 124 general-risk-tolerance lead SNPs are in weak LD (pairwise $r^2 > 0.1$) with a lead SNP of at least one of the other GWAS (including 45 for adventurousness and 49 for at least one of the four risky behaviors or their first PC). To empirically assess if this overlap could be attributed to chance, we conducted resampling exercises under the null hypothesis that the lead SNPs of our supplementary GWAS are distributed independently of the general-risk-tolerance loci and lead SNPs. We strongly rejected this null hypothesis ($P < 0.0001$; **Supplementary Note**).

Several long-range LD regions, candidate inversions, and LD blocks[19] stand out for being associated both with general risk tolerance and with all or most of the supplementary phenotypes. We tested whether the signs of the lead SNPs located in these regions tend to be concordant across our primary and supplementary GWAS. We strongly rejected the null hypothesis of no concordance ($P < 3\times10^{-30}$; **Supplementary Note**), suggesting that these regions represent shared genetic influences, rather than colocalization of causal SNPs. **Figure 2.1b** and **Supplementary Fig. 6** show local Manhattan plots for some of these long-range LD regions and candidate inversions. The long-range LD region[18] on chromosome 3 (~83.4 to 86.9 Mb) contains lead SNPs from all seven GWAS as well as the most significant lead SNP from the general-risk-tolerance GWAS, rs993137 ($P = 2.14\times10^{-40}$), which is located in the gene *CADM2*. Another long-range LD region, on chromosome 6 (~25.3 to 33.4 Mb), covers the HLA-complex and contains lead SNPs from all GWAS except drinks per week. Three candidate inversions on chromosomes 7 (~124.6 to 132.7 Mb), 8 (~7.89 to 11.8 Mb), and 18 (~49.1 to 55.5 Mb) contain lead SNPs from six, five, and all seven of our GWAS, respectively. Finally, four other LD blocks[19] that do not overlap known long-range LD or candidate inversion regions each contain lead SNPs from five of our GWAS (including general risk tolerance). While many of the lead SNPs in these regions are not conditional associations, the above results regarding the numbers of GWAS with lead SNPs in these regions also hold if we only consider the conditional associations instead of the lead SNPs in those regions. The two long-range LD regions and the three candidate inversions have previously been found to be associated with numerous phenotypes, including many cognitive and neuropsychiatric phenotypes[20].

To investigate genetic overlap at the genome-wide level, we estimated genetic correlations with self-reported general risk tolerance using bivariate LD Score regression[9]. (For this and all subsequent analyses involving general risk tolerance, we used the summary statistics from the combined meta-analysis of our discovery and replication GWAS.) The estimated genetic correlations with our six supplementary phenotypes are all positive, larger than ~0.25, and highly significant ($P < 2.3\times10^{-30}$; **Figure 2.2**), indicating that SNPs associated with higher general risk tolerance also tend to be associated with riskier behavior. The largest estimated genetic correlations are with adventurousness ($\hat{r}_g = 0.83$, $SE = 0.01$), number of sexual partners (0.52, $SE = 0.02$), automobile speeding propensity (0.45, $SE = 0.02$), and the first PC of the four risky behaviors (0.50, $SE = 0.02$).

Our estimates of the genetic correlations between general risk tolerance and the supplementary risky behaviors are substantially higher than the corresponding phenotypic correlations (**Supplementary Tables 8** and **9**). Although measurement error partly accounts for the low phenotypic correlations, the genetic correlations remain considerably higher even after adjustment of the phenotypic correlations for measurement error. The comparatively large genetic correlations support the view that a general factor of risk tolerance partly accounts for cross-domain correlation in risky behavior[21,22] and imply that this factor is genetically influenced. The lower phenotypic correlations suggest that environmental factors are more important contributors to domain-specific risky behavior[23,24].

To increase the precision of our estimates of the SNPs' effects on general risk tolerance, we leveraged the high degree of genetic overlap across our phenotypes by conducting Multi-Trait Analysis of GWAS (MTAG)[16]. We used as inputs the summary statistics of our GWAS of general risk tolerance, of our first five supplementary GWAS (i.e., not including the first PC of the four risky behaviors), and of a previously published GWAS on lifetime cannabis use[25] (**Supplementary Note**). MTAG increased the number of general-risk-tolerance lead SNPs from 124 to 312 (**Supplementary Fig. 7** and **Supplementary Table 10**).

We also estimated genetic correlations between general risk tolerance and 28 additional phenotypes (**Figure 2.2** and in **Supplementary Table 9**). These included phenotypes for which we could obtain summary statistics from previous GWAS, as well as five phenotypes for which we conducted new GWAS. The estimated genetic correlations for the personality traits extraversion ($\hat{r}_g$ = 0.51, *SE* = 0.03), neuroticism (–0.42, *SE* = 0.04), and openness to experience (0.33, *SE* = 0.03) are significantly distinguishable from zero after Bonferroni correction and are substantially larger in magnitude than previously reported phenotypic correlations[26], pointing to shared genetic influences among general risk tolerance and these traits. After Bonferroni correction, we also found significant positive genetic correlations with the neuropsychiatric phenotypes ADHD, bipolar disorder, and schizophrenia. Viewed in light of the genetic correlations we found with some supplementary phenotypes and additional risky behaviors classified as externalizing (e.g., substance use, elevated sexual behavior, and fast driving), these results suggest the hypothesis that the overlap with the neuropsychiatric phenotypes is driven by their externalizing component[27].

Polygenic prediction

We constructed polygenic scores of general risk tolerance to gauge their potential usefulness in empirical research (**Supplementary Note**). We used the Add Health, HRS, NTR, STR, UKB-siblings, and Zurich cohorts as validation cohorts (**Supplementary Table 5** provides an overview of these cohorts; the UKB-siblings cohort comprised individuals with at least one full sibling in the UKB). For each validation cohort, we constructed the score using summary statistics from a meta-analysis of our discovery and replication GWAS that excluded the cohort (for the UKB-siblings cohort, we reran our UKB GWAS after excluding individuals from that cohort). Our measure of predictive power is the incremental R^2 (or pseudo-R^2) from adding the score to a regression of the phenotype on controls for sex, birth year, and the top ten principal components of the genetic relatedness matrix.

Our preferred score was constructed with LDpred[28]. Our largest validation cohort ($n \sim 35{,}000$) is the UKB-siblings cohort. In that validation cohort, the score's predictive power is 1.6% for general risk tolerance, 1.0% for the first PC of the four risky behaviors, 0.8% for number of sexual partners, 0.6% for automobile speeding propensity, and ~0.15% for drinks per week and ever smoker. Across our validation cohorts, in which other phenotypes are measured, the score is also predictive of several personality phenotypes and a suite of real-world measures of risky behaviors in the health, financial, career, and other domains (**Supplementary Figs. 8-9** and **Supplementary Tables 11-14**). The incremental R^2 we observe for general risk tolerance is consistent with our theoretical prediction, given the GWAS sample sizes, the SNP heritability of general risk tolerance (**Table 2.1**), and the imperfect genetic correlations across the GWAS and validation cohorts[29,30] (**Supplementary Note**).

Biological annotation

To gain insights into the biological mechanisms through which genetic variation influences general risk tolerance, we conducted a number of bioinformatics analyses using the results of the combined meta-analysis of our discovery and replication GWAS of general risk tolerance.

First, we systematically reviewed the literature that aimed to link risk tolerance to biological pathways (**Supplementary Note**). Our review covered studies based on candidate genes (i.e., specific genetic variants used as proxies for biological pathways), pharmacological manipulations, biochemical assays, genetic manipulations in rodents, as well as other research designs. Our review identified 132 articles that matched our search criteria (**Supplementary Table 2.1**). This previous work has focused on five main biological pathways: the steroid hormone cortisol, the monoamines dopamine and serotonin, and the steroid sex hormones estrogen and testosterone. Using a MAGMA[31] competitive gene-set analysis, we found no evidence that SNPs within genes associated with these five pathways tend to be more associated with general risk tolerance than SNPs in other genes (**Supplementary Table 15**). Furthermore, none of the other bioinformatics analyses we report below point to these pathways.

We also examined the 15 most commonly tested autosomal genes within the dopamine and serotonin pathways, which were the focus of most of the 34 candidate-gene studies identified by our literature review. We verified that the SNPs available in our GWAS results tag most of the genetic variants typically used to test the 15 genes. Across one SNP-based test and two gene-based tests, we found no evidence of non-negligible associations between those genes and general risk tolerance (**Figure 2.1c** and **Supplementary Table 16**). (We note, however, that some brain regions identified in analyses we report below are areas where dopamine and serotonin play important roles.)

Second, we performed a MAGMA[31] gene analysis to test each of ~18,000 protein-coding genes for association with general risk tolerance (**Supplementary Note**). After Bonferroni correction, 285 genes were significant (**Supplementary Fig. 10** and **Supplementary Table 17**). To gain insight into the functions and expression patterns of these 285 genes, we looked them up in the Gene Network[32] co-expression database.

Third, to identify relevant biological pathways and identify tissues in which genes near general-risk-tolerance-associated SNPs are expressed, we applied the software tool DEPICT[33] to the SNPs with *P* values less than 10^{-5} in our GWAS of general risk tolerance (**Supplementary Note**).

Both the Gene Network and the DEPICT analyses separately point to a role for glutamate and GABA neurotransmitters, which are the main excitatory and inhibitory neurotransmitters in the brain, respectively[34] (**Figure 2.3a** and **Supplementary Tables 18** and **19**). To our knowledge, with the exception of a recent study[35] prioritizing a much larger number of genes and pathways, no published large-scale GWAS of cognition, personality, or neuropsychiatric phenotypes has pointed to clear roles both for glutamate and GABA (although glutamatergic neurotransmission has been implicated in recent GWAS of schizophrenia[36] and major depression[37]). Our results suggest that the balance between excitatory and inhibitory neurotransmission may contribute to variation in general risk tolerance across individuals.

The Gene Network and the DEPICT tissue enrichment analyses also both separately point to enrichment of the prefrontal cortex and the basal ganglia (**Figure 2.3b** and **Supplementary Tables 18**, **20**, and **21**). The cortical and subcortical regions highlighted by DEPICT include some of the major components of the cortical-basal ganglia circuit, which is known as the reward system in human and non-human primates and is critically involved in learning, motivation, and decision-making, notably under risk and uncertainty[38,39]. We caution, however, that our results do not point exclusively to the reward system.

Lastly, we used stratified LD Score regression[40] to test for the enrichment of SNPs associated with histone marks in 10 tissue or cell types (**Supplementary Note**). Central nervous system tissues are the most enriched, accounting for 44% (*SE* = 3%) of the heritability while comprising only 15% of the SNPs (**Supplementary Fig. 11a** and **Supplementary Table 22**). Immune/hematopoietic tissues are also significantly enriched. While a role for the immune system in modulating risk tolerance is plausible given prior evidence of its involvement in several neuropsychiatric disorders[36,37], future work is needed to confirm this result and to uncover specific pathways that might be involved.

Discussion

Our results provide insights into biological mechanisms that influence general risk tolerance. Our bioinformatics analyses point to the role of gene expression in brain regions that have been identified by neuroscientific studies on decision-making, notably the prefrontal cortex, basal ganglia, and midbrain, thereby providing convergent evidence with that from neuroscience[38,39]. Yet our analyses failed to find evidence for the main biological pathways that had been previously hypothesized to influence risk tolerance. Instead, our analyses implicate genes involved in glutamatergic and GABAergic neurotransmission, which were heretofore not generally believed to play a noteworthy role in risk tolerance.

Although our focus has been on the genetics of general risk tolerance and risky behaviors, environmental and demographic factors account for a substantial share of these phenotypes' variation. We observe sizeable effects of sex and age on general risk tolerance in the UKB data (**Supplementary Fig. 4**), and life experiences have been shown to affect both measured risk tolerance and risky behaviors (e.g., refs. [41,42]). The GWAS results we have generated will allow researchers to construct and use polygenic scores of general risk tolerance to measure how environmental, demographic, and genetic factors interact with one another.

For the behavioral sciences, our results bear on an ongoing debate about the extent to which risk tolerance is a "domain-general" as opposed to a "domain-specific" trait. Low phenotypic correlations in risk tolerance across decision-making domains have been interpreted as supporting the domain-specific view[23,24]. Across the risky behaviors we study, we found that the genetic correlations were considerably higher than the phenotypic correlations (even after the latter are corrected for measurement error) and that many lead SNPs are shared across our phenotypes. These observations suggest that the low phenotypic correlations across domains are due to environmental factors that dilute the effects of a genetically-influenced domain-general factor of risk tolerance.

Authors – Chapter 2

Richard Karlsson Linnér, Pietro Biroli, Edward Kong, S Fleur W Meddens, Robbee Wedow, Mark Alan Fontana, Maël Lebreton, Stephen P Tino, Abdel Abdellaoui, Anke R Hammerschlag, Michel G Nivard, Aysu Okbay, Cornelius A Rietveld, Pascal N Timshel, Maciej Trzaskowski, Ronald de Vlaming, Christian L Zünd, Yanchun Bao, Laura Buzdugan, Ann H Caplin, Chia-Yen Chen, Peter Eibich, Pierre Fontanillas, Juan R Gonzalez, Peter K Joshi, Ville Karhunen, Aaron Kleinman, Remy Z Levin, Christina M Lill, Gerardus A Meddens, Gerard Muntané, Sandra Sanchez-Roige, Frank J van Rooij, Erdogan Taskesen, Yang Wu, Futao Zhang, *23andMe Research Team*, *eQTLgen Consortium*, *International Cannabis Consortium*, *Social Science Genetic Association Consortium*, Adam Auton, Jason D Boardman, David W Clark, Andrew Conlin, Conor C Dolan, Urs Fischbacher, Patrick J F Groenen, Kathleen Mullan Harris, Gregor Hasler, Albert Hofman, Mohammad A Ikram, Sonia Jain, Robert Karlsson, Ronald C Kessler, Maarten Kooyman, James MacKillop, Minna Männikkö, Carlos Morcillo-Suarez, Matthew B McQueen, Klaus M Schmidt, Melissa C Smart, Matthias Sutter, A Roy Thurik, André G Uitterlinden, Jon White, Harriet de Wit, Jian Yang, Lars Bertram, Dorret I Boomsma, Tõnu Esko, Ernst Fehr, David A Hinds, Magnus Johannesson, Meena Kumari, David Laibson, Patrik K E Magnusson, Michelle N Meyer, Arcadi Navarro, Abraham A Palmer, Tune H Pers, Danielle Posthuma, Daniel Schunk, Murray B Stein, Rauli Svento, Henning Tiemeier, Paul R H J Timmers, Patrick Turley, Robert J Ursano, Gert G Wagner, James F Wilson, Jacob Gratten, James J Lee, David Cesarini, Daniel J Benjamin, Philipp D Koellinger, and Jonathan P Beauchamp

References – Chapter 2

1. Dohmen, T. *et al.* Individual risk attitudes: Measurement, determinants, and behavioral consequences. *J. Eur. Econ. Assoc.* **9,** 522–550 (2011).
2. Falk, A., Dohmen, T., Falk, A. & Huffman, D. The nature and predictive power of preferences: Global evidence. *IZA Discuss. Pap.* (2015).
3. Beauchamp, J. P., Cesarini, D. & Johannesson, M. The psychometric and empirical properties of measures of risk preferences. *J. Risk Uncertain.* **54,** 203–237 (2017).
4. Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P. & Wallace, B. Genetic variation in preferences for giving and risk taking. *Q. J. Econ.* **124,** 809–842 (2009).
5. Harden, K. P. *et al.* Beyond dual systems: A genetically-informed, latent factor model of behavioral and self-report measures related to adolescent risk-taking. *Dev. Cogn. Neurosci.* **25,** 221–234 (2017).
6. Hewitt, J. K. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behav. Genet.* **42,** 1–2 (2012).
7. Day, F. R. *et al.* Physical and neurobehavioral determinants of reproductive onset and success. *Nat. Genet.* **48,** 617–623 (2016).
8. Strawbridge, R. J. *et al.* Genome-wide analysis of self-reported risk-taking behaviour and cross-disorder genetic correlations in the UK Biobank cohort. *Transl. Psychiatry* **8,** 1–11 (2018).
9. Bulik-Sullivan, B. K. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).
10. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624–633 (2016).
11. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
12. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
13. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–375 (2012).
14. Byrnes, J. P., Miller, D. C. & Schafer, W. D. Gender differences in risk taking: a meta-analysis. *Psychol. Bull.* **125,** 367–383 (1999).
15. Croson, R. & Gneezy, U. Gender Differences in Preferences. *J. Econ. Lit.* **47,** 448–474

(2009).

16. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50,** 229–237 (2018).

17. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *bioRxiv* **99,** 1–28 (2016).

18. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83,** 132–139 (2008).

19. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32,** 283–285 (2016).

20. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-1006 (2014).

21. Einav, B. L., Finkelstein, A., Pascu, I. & Cullen, M. R. How general are risk preferences? Choices under uncertainty in different domains. *Am. Econ. Rev.* **102,** 2606–2638 (2016).

22. Frey, R., Pedroni, A., Mata, R., Rieskamp, J. & Hertwig, R. Risk preference shares the psychometric structure of major psychological traits. *Sci. Adv.* **3,** e1701381 (2017).

23. Weber, E. U., Blais, A. E. & Betz, N. E. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *J. Behav. Decis. Mak. J. Behav. Dec. Mak.* **15,** 263–290 (2002).

24. Hanoch, Y., Johnson, J. G. & Wilke, A. Domain specificity in experimental measures and participant recruitment: an application to risk-taking behavior. *Psychol. Sci.* **17,** 300–304 (2006).

25. Stringer, S. *et al.* Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32,330 subjects from the International Cannabis Consortium. *Transl. Psychiatry* **6,** e769 (2016).

26. Becker, A., Deckers, T., Dohmen, T., Falk, A. & Kosse, F. The relationship between economic preferences and psychological personality measures. *Annu. Rev. Econom.* **4,** 453–478 (2012).

27. Krueger, R. F. *et al.* Etiologic connections among substance dependence, antisocial behavior and personality: Modeling the externalizing spectrum. *J. Abnorm. Psychol.* **111,** 411–424 (2002).

28. Vilhjálmsson, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).

29. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3,** e3395 (2008).

30. de Vlaming, R. *et al.* Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLoS Genet.* **13,** e1006495 (2017).
31. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11,** 1–19 (2015).
32. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).
33. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).
34. Petroff, O. A. C. GABA and glutamate in the human brain. *Neurosci.* **8,** 562–573 (2002).
35. Lee, J. *et al.* Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nat. Genet.* **50,** 1112–1121 (2018).
36. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).
37. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv* (2017). doi:https://doi.org/10.1101/167577
38. Haber, S. N. & Knutson, B. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* **35,** 4–26 (2010).
39. Tobler, P. N. & Weber, E. U. in *Neuroeconomics* 149–172 (Elsevier, 2014). doi:10.1016/B978-0-12-416008-8.00009-7
40. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).
41. Sahm, C. R. How much does risk tolerance change? *Q. J. Financ.* **2,** 1250020 (2012).
42. Malmendier, U. & Nagel, S. Depression babies: Do macroeconomic experiences affect risk taking? *Q. J. Econ.* **126,** 373–416 (2011).
43. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315,** (2007).

Figures and tables – Chapter 2

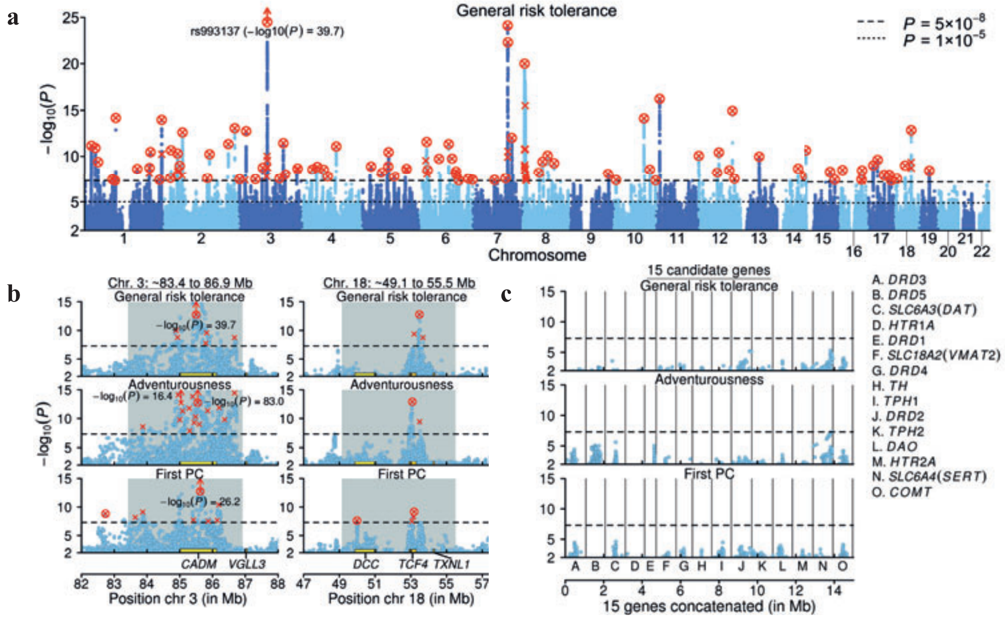

Figure 2.1. Manhattan plots. In all panels, the *x*-axis is chromosomal position; the *y*-axis is the GWAS *P* value on a $-\log_{10}$ scale (based on a two-tailed *z*-test); each lead SNP is marked by a red “×”; each conditional association is marked by a red “○”; and each SNP that is both a lead SNP and a conditional association is marked by a red “⊗”. **a**, Manhattan plots for the discovery GWAS of general risk tolerance (*n* = 939,908). **b**, Local Manhattan plots of a long-range LD region on chromosome 3 and a candidate inversion on chromosome 18 that contain lead SNPs for all seven of our GWAS. The gray background marks the locations of long-range LD or candidate inversion regions. **c**, Local Manhattan plots of the areas around the 15 most commonly tested candidate genes in the prior literature on the genetics of risk tolerance. Each local plot shows all SNPs within 500 kb of the gene’s borders that are in weak LD ($r^2 > 0.1$) with a SNP in the gene. The 15 plots are concatenated and shown together in the panel, divided by the black vertical lines. The 15 genes are not particularly strongly associated with general risk tolerance or the risky behaviors, as can be seen by comparing the results within each row across panels **b** and **c** (the three rows correspond to the GWAS of general risk tolerance, adventurousness (*n* = 557,923), and the first PC of the four risky behaviors (*n* = 315,894)).

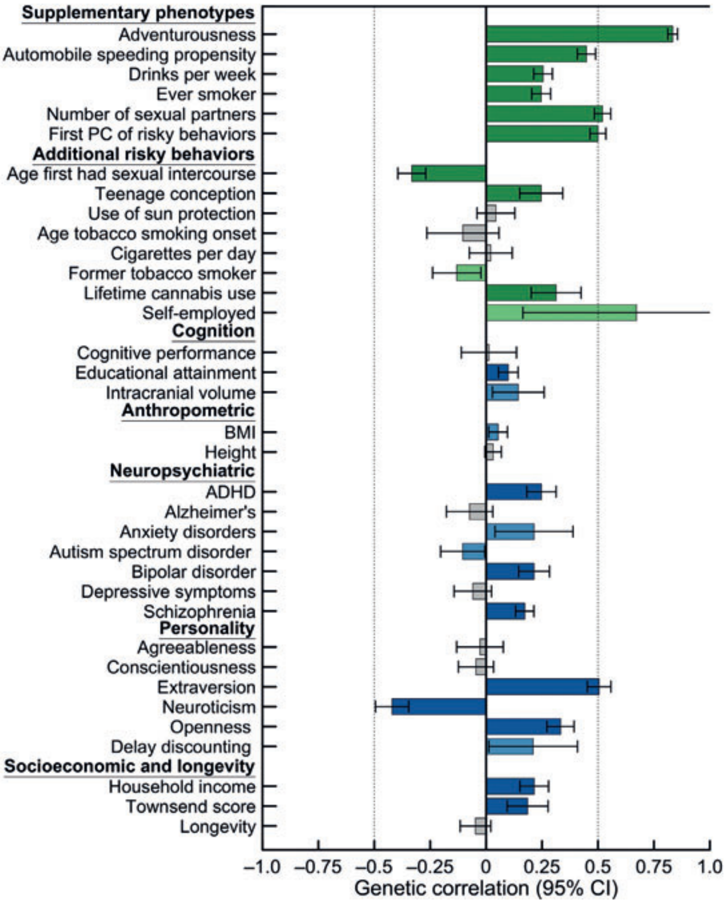

Figure 2.2. Genetic correlations with general risk tolerance. The genetic correlations were estimated using bivariate LD Score (LDSC) regression[9]. Error bars show 95% confidence intervals. For the supplementary phenotypes and the additional risky behaviors, green bars represent significant estimates with the expected signs, where higher risk tolerance is associated with riskier behavior. For the other phenotypes, blue bars represent significant estimates. Light green and light blue bars represent genetic correlations that are statistically significant at the 5% level, and dark green and dark blue bars represent correlations that are statistically significant after Bonferroni correction for 35 tests (the total number of phenotypes tested). Grey bars represent correlations that are not statistically significant at the 5% level. The two dotted vertical lines indicate genetic correlations of –0.5 and 0.5, respectively. All significance tests are two-sided.

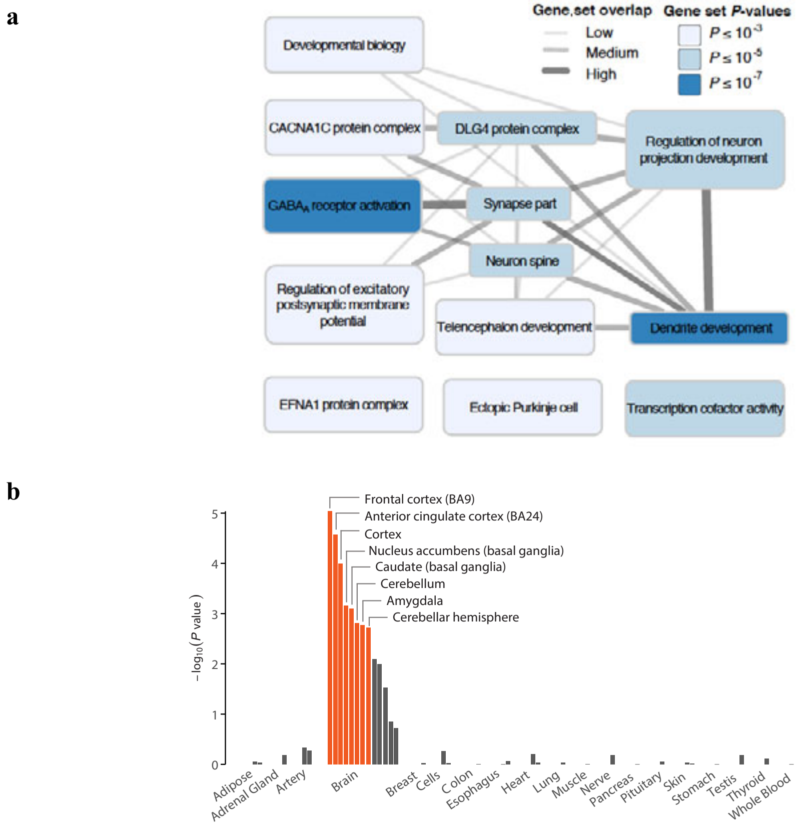

Figure 2.3. Results from selected biological analyses. a, DEPICT gene-set enrichment diagram. We identified 93 reconstituted gene sets that are significantly enriched (FDR < 0.01) for genes overlapping DEPICT-defined loci associated with general risk tolerance; using the Affinity Propagation method[43], these were grouped into the 13 clusters displayed in the graph. Each cluster was named after its exemplary gene set, as chosen by the Affinity Propagation tool, and each cluster's color represents the permutation *P* value of its most significant gene set. The "synapse part" cluster includes the gene set "glutamate receptor activity," and several members of the "GABA$_A$ receptor activation" cluster are defined by gamma-aminobutyric acid signaling. Overlap between the named representatives of two clusters is represented by an edge. Edge width represents the Pearson correlation ρ between the two respective vectors of gene membership scores ($\rho < 0.3$, no edge; $0.3 \leq \rho < 0.5$, thin edge; $0.5 \leq \rho < 0.7$, intermediate edge; $\rho \geq 0.7$, thick edge). **b**, Results of DEPICT tissue enrichment analysis using GTEx data. The panel shows whether the genes overlapping DEPICT-defined loci associated with general risk tolerance are significantly overexpressed (relative to genes in random sets of loci matched by gene density) in various tissues. Tissues are grouped by organ or tissue type. The orange bars correspond to tissues with significant overexpression (FDR < 0.01). The *y*-axis is the significance on a $-\log_{10}$ scale. See **Supplementary Note** for additional details.

Table 2.1 | GWAS results

GWAS	Cohorts analyzed	*n*	Mean χ^2	LD Score intercept (*SE*)	# lead SNPs	# loci	# cond. assoc.	SNP h^2 (*SE*)
General risk tolerance (disc. GWAS)	UKB; 23andMe	939,908	1.85	1.04 (0.01)	124	99	91	0.046 (0.001)
General risk tolerance (repl. GWAS)	10 indep. cohorts	35,445	1.03	1.00 (0.07)	0	0	0	--
General risk tolerance (disc. + repl.)	UKB; 23andMe; 10 indep. cohorts	975,353	1.87	1.04 (0.01)	132	107	97	0.045 (0.001)
Adventurousness	23andMe	557,923	1.98	1.05 (0.01)	167	137	126	0.098 (0.002)
Automobile speeding propensity	UKB	404,291	1.53	1.03 (0.01)	42	36	33	0.079 (0.003)
Drinks per week	UKB	414,343	1.61	1.03 (0.01)	85	62	61	0.085 (0.003)
Ever smoker	UKB; TAG Consortium[44]	518,633	1.97	1.05 (0.01)	223	183	172	0.109 (0.003)
Number of sexual partners	UKB	370,711	1.77	1.04 (0.01)	117	97	88	0.128 (0.003)
First PC of the four risky behaviors	UKB	315,894	1.77	1.05 (0.01)	106	89	84	0.156 (0.004)

The table provides an overview of the GWAS of our primary and supplementary phenotypes. Replication analysis of the lead SNPs' association results in independent cohorts was only conducted for the discovery GWAS of general risk tolerance. "*n*": GWAS sample size; "Mean χ^2": mean GWAS chi-squared statistics across HapMap3 SNPs with minor allele frequency (MAF) greater than 0.01; "LD Score intercept": estimate of the intercept from a LD Score regression[11] using HapMap3 SNPs with MAF greater than 0.01; "# lead SNPs": number of approximately independent (pairwise $r^2 < 0.1$) lead SNPs; "# loci": number of associated loci; "# cond. assoc.": number of conditional associations in the COJO analysis[13]; "SNP h^2": SNP heritability estimated with the Heritability Estimator from Summary Statistics (HESS) method[17] using 1000 Genomes phase 3 SNPs with MAF greater than 0.05; "disc.": discovery; "repl.": replication; "indep.": independent.

Chapter 2 Supplementary Methods

Overview of main GWAS

All analyses were performed at the cohort level according to a pre-specified and publicly archived analysis plan[1]. The original analysis plan was archived on February 4, 2016. For self-reported general risk tolerance, the analysis plan specified that the discovery GWAS would be conducted in the UKB and that the replication would be carried out in a meta-analysis of all other cohorts.

We updated the analysis plan on November 9, 2016 to include the analysis of the four risky behaviors and their first PC. For these phenotypes, the analysis plan specified that the GWAS would be conducted in the UKB. We did not attempt replication for these phenotypes. We updated the analysis plan a second time on July 10, 2017 to add the 23andMe cohort to the discovery GWAS of self-reported general risk tolerance. Two minor updates were made on August 7 and September 14, 2017 to specify that follow-up analyses (such as polygenic prediction) would be performed, whenever possible, on a meta-analysis combining the discovery and replication GWAS of general risk tolerance, and that we would add a GWAS of adventurousness alongside the primary GWAS of general risk tolerance and the other supplementary GWAS.

Cohorts other than the UKB could join this study by first supplying descriptive statistics and thereafter GWAS summary statistics from GWAS of self-reported general risk tolerance in November 2015 and December 2015, respectively. Two additional cohorts, Army STARRS and VHSS, joined the study later and provided descriptive and GWAS summary statistics in late 2016. Summary statistics were uploaded to a central, secure server and subsequently meta-analyzed. The lead PI of each cohort affirmed that the results contributed to the study were based on analyses approved by the local Research Ethics Committee and/or Institutional Review Board responsible for overseeing research. All participants provided written informed consent. We also obtained the descriptive and GWAS summary statistics from GWAS of self-reported general risk tolerance and adventurousness from 23andMe in late July 2017. An overview of the participating cohorts is reported in **Supplementary Table 5**.

The analysis plan instructed all cohorts to limit the analysis to individuals of European ancestry, to exclude individuals with missing covariates, to remove samples that displayed a SNP call rate of less than 95%, and to apply cohort-specific standard quality control filters before imputation. The cohort-specific standard quality control filters are reported in **Supplementary Table 24**. GWAS were limited to the 22 autosomes. The cohorts were required to provide unfiltered GWAS summary statistics including the following information for each SNP: chromosome and base-pair position, rsID, effect-coded allele, other allele, sample size per SNP, coefficient estimate (beta), standard error of the coefficient estimate, *P* value of the association uncorrected for genomic control, effect-coded allele frequency (EAF), imputation status, imputation quality, and Hardy-Weinberg equilibrium exact test *P* value for directly genotyped markers.

The analysis plan included power calculations assuming that 100,000 individuals in the UKB answered “Yes” (“cases”) to the general-risk-tolerance question, and 270,000 individuals answered “No” (“controls”). Under this assumption, our study would have 73% power to detect single nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) of 0.3 and an odds ratio of 1.05 with a genome-wide significance threshold of $P = 5\times10^{-8}$.

For general risk tolerance, the final sample size for the “discovery GWAS” meta-analysis of the UKB and 23andMe cohorts was 939,908 individuals. Replication was performed in a meta-analysis of 10 independent cohorts from seven studies totaling 35,445 individuals. We will henceforth refer to this meta-analysis of the 10 replication cohorts as the “replication GWAS.” The follow-up analyses we describe in the following **Supplementary Note sections** were performed with GWAS summary statistics from a meta-analysis combining the discovery and replication GWAS (n = 975,353), except where otherwise noted.

For adventurousness, GWAS summary statistics from 23andMe were analyzed (n = 557,923).

For three of the four risky behaviors, namely automobile speeding propensity (n = 404,291), drinks per week (n = 414,343), and number of sexual partners (n = 370,711), and for the first PC of the four risky behaviors (n = 315,894), GWAS were conducted in the UKB only, as specified in the updated analysis plan. For the remaining risky behavior, ever smoker, we meta-analyzed the summary statistics from the UKB GWAS (n = 444,598) with those from the Tobacco, Alcohol and Genetics (TAG) Consortium[2] (n = 74,035), leading to a total sample size of 518,633 (the TAG consortium refers to the ever smoker phenotype as "smoking initiation").

Genotyping and imputation

Genotyping[d] was performed using a range of common, commercially available genotyping arrays. An overview of the genotyping and imputation procedure is provided in **Supplementary Table 24**. The participating cohorts were encouraged to use their standard quality-control protocols before imputation, as long as the applied filters satisfied the minimum requirements specified in the analysis plan (SNP call rate > 95%, HWE exact test P value > 10^{-6}, MAF > 1%). For the UKB, different filters were used, following ref.[3].

The cohorts, except for 23andMe, Army STARRS, BASE-II, UKB, and VHSS imputed markers using the 1000 Genomes phase 1 reference panel (March 2012 release version 3). 23andMe used the 1000 Genomes phase 1 (September 2013 haplotype release). Army STARSS used the 1000 Genomes phase 1 (August 2012 haplotype release). BASE-II used the more recent reference panel 1000 Genomes phase 3 (October 2014 haplotype release version 5). UKB used a customized reference panel based on the Haplotype Reference Consortium release 1.1 combined with the UK10K haplotype reference panel[4]. VHSS used the Haplotype Reference Consortium release 1.1[5].

All genetic positions reported in this study are denoted with those of the Genome Reference Consortium’s human assembly 37 (GRCh37, sometimes referred to as the National Center for Biotechnology Information hg19).

Association analyses

Cohorts were encouraged to exclude individuals with SNP call rates less than 95%, with excessive autosomal heterozygosity, and with sex mismatch. Family-based cohorts were informed to control for family structure either with mixed linear modeling or with a procedure selecting only one individual in each pair that displayed relatedness greater than 5% in a genetic relatedness matrix.

The genome-wide association analysis performed in each cohort estimated the following regression for each SNP, as in Okbay *et al.* 2016[6]:

[d] The UKB genotype data was handled with QCtool, available at http://www.well.ox.ac.uk/~gav/qctool/#overview

$$Y_i = \beta_0 + \beta_1 SNP_i + \boldsymbol{PC_i}\boldsymbol{\gamma} + \boldsymbol{X_i}\boldsymbol{\alpha} + \boldsymbol{C_i}\boldsymbol{\theta} + \epsilon_i, \quad (1)$$

where Y_i is the phenotype for individual i, SNP_i is the number of effect-coded alleles of the SNP[e], $\boldsymbol{PC_i}$ is a vector of principal components of the genetic relatedness matrix after application of the pre-imputation filters described above, and $\boldsymbol{X_i}$ is a vector of control variables. In all cohorts, $\boldsymbol{X}$ included controls for sex and birth year. In most cohorts (including the 23andMe cohort), these included sex, birth year, birth year squared, birth year cubed, and the interactions between sex and the three birth-year variables; in the UKB, these included sex-specific birth year fixed effects. $\boldsymbol{C_i}$ is a vector containing cohort-specific controls and technical covariates (such as dummy variables for genotyping array and genotyping batches) that are recommended in the analysis plan. All associations were performed with males and females pooled. A summary of the GWAS association models and control variables for each cohort is reported in **Supplementary Table 2**.

In practice, the phenotype was often residualized by first regressing the phenotype on the control variables, and the residualized phenotype was then regressed on the genotypes. This approach leads to almost identical results as estimating the full regression model directly while drastically reducing the computation time needed for the GWAS.

Linear mixed models in the UKB

The association analyses in the UKB were performed with linear mixed models (LMM) with the BOLT-LMM v2.2 software[7]. The benefit of LMM is that the method accounts for cryptic relatedness and population structure, which allows the inclusion of related individuals in the sample, thereby yielding a larger sample size and greater statistical power. Using LMM is computationally intensive, and the BOLT algorithm is a new method that makes LMM analysis of hundreds of thousands of individuals computationally feasible. The method requires a set of SNPs to be included in the genetic variance component, and we included 483,680 directly genotyped bi-allelic autosomal SNPs with MAF > 0.005 and HWE P value > 10^{-6}. We included individuals based on self-reported ancestry, specifically those who self-reported to be of "white" ancestry (i.e., self-reported white, British, Irish, or any other white background). In addition, we limited the GWAS to individuals for whom the value of the first principal component of the genetic relatedness matrix was less than "0," which identifies the cluster of individuals of European ancestry. We dropped individuals whose reported sex did not match their genetic sex, individuals with putative sex chromosome aneuploidy, individuals that did not pass the UKB internal genotype quality control, and individuals with missing values.

Main reference panel

The full release of the UK Biobank genetic data was imputed with haplotypes from the Haplotype Reference Consortium v1.1 (HRC) and the UK10K haplotype reference panel[4]. A recommendation was communicated soon after the release of the genotype data in July 2017. It was recommended that only SNPs available in the HRC be used for analysis, because a subset of variants imputed from the UK10K reference panel have wrongfully imputed genomic positions, while none of the HRC SNPs are affected. We therefore used the HRC v.1.1 as the reference panel for quality control of the GWAS summary statistics, and to determine the independence of significant SNPs. The following section describes our quality control of the

[e] For imputed SNPs we used best-guess data for samples that were imputed with IMPUTE[46], and dosage data for samples that were imputed with MaCH/Minimac[47].

HRC whole-genome sequence data (WGS) when constructing the reference panel. We will hereafter refer to the resulting reference panel as the "main reference panel."

Quality-control of the main reference panel

The HRC haplotypes were downloaded from the European Genome-phenome Archive (EGA) on August 1, 2017. Strict internal quality control had already been applied to the WGS data, as described in-depth elsewhere[5], and we restricted the main reference panel to variants that passed all pre-applied genotype call filters (i.e., variants whose VCF FILTER status is "PASS"; these pre-applied filters included, among other filters, a filter to remove variants with minor allele count (MAC) ≤ 5.) Before our internal QC the WGS data contained 40,405,506 autosomal and X-chromosome SNPs. The following protocol was restricted to the 39,131,579 autosomal SNPs because the pre-registered analysis plan restricts our analyses to the autosomes. The HRC reference panel does not include any structural variants, such as INDELs[5].

We performed a series of best-practice alignments of the WGS data for consistent and unique identification of variants[8] with the open-source software BCFtools[f] created by the Wellcome Trust Sanger Institute. Because PLINK cannot properly handle truly multi-allelic variants, we split multi-allelic variants into multiple bi-allelic variants. We then confirmed that all reference alleles and genomic positions matched the Genome Reference Consortium Human genome build 37 (GRCh37)[9]. To avoid issues with chromosomal positions mapping to multiple NCBI marker IDs (rsIDs), all rsIDs were removed, and all variants were given a unique identifier (henceforth referred to the as the "unique ID") in the form of chromosome, base-pair position, reference allele, and alternative allele, separated by colons (e.g., 1:123456:C:T). Using this format for variant IDs, together with the alignment to the reference genome GRCh37, ensures a unique representation of all SNPs and a lack of duplicate variants with switched reference alleles (e.g., 1:123456:C:T and 1:123456:T:C).

To avoid possible issues with inconsistencies with the UK10K haplotype reference in future work that may use that reference, we investigated possible strand and allele issues across the reference panels. By comparing the reference alleles and allele frequencies we found 24,394 variants with inconsistent alleles, and we decided to drop these from the main reference panel so that they would be removed during QC of the GWAS summary statistics.

We converted the VCF data to PLINK binary format with PLINK v.1.9b3.46[10], and we removed all multi-allelic variants (without retaining any of the multi-allelic variants coded as bi-allelic SNPs). Monomorphic SNPs (i.e., SNPs with MAF = 0) were kept in the reference panel as recommended[5]. We thereafter excluded one member of each pair of individuals with genomic relatedness greater than 0.025 from the sample, which removed 4,917 individuals of the 22,691 individuals for whom data were available for all autosomes in the VCF data.

In summary, the main reference panel consists of 17,774 individuals and includes 38,889,224 bi-allelic autosomal SNPs that passed QC.

rsID mapping

Since rsIDs were removed from the SNPs to ensure unique identification, we created a map file so that each SNP could be assigned an rsID after meta-analysis, which is performed on the unique ID format described above (e.g., 1:123456:C:T). The map file contains the rsIDs from the HRC v1.1 sites list, and because we removed all multi-allelic variants there are no duplicate rsIDs.

[f] BCFtools can be downloaded here: http://samtools.github.io/bcftools/bcftools.html

UK Biobank genotyping arrays

Combining data from the UK BiLEVE and the UK Biobank Axiom arrays

The participants of the UKB were genotyped with two different but similar genotyping arrays[4,11–13]. UKB participants who were enrolled in the UK BiLEVE study (a study of smoking behavior, lung function, and chronic obstructive pulmonary disease[11]) were genotyped with the UK BiLEVE array (*n* ~ 50,000), and the remaining participants were genotyped with the UK Biobank Axiom array (*n* ~ 400,000). Henceforth, we will refer to these two sets of UK Biobank participants as the UK BiLEVE and the UKB Axiom cohorts.

While the UK Biobank (UKB) is a population-based study[12] collected during 2006–2010, the sample selection for the UK BiLEVE study began at a later stage, in 2012. The UK BiLEVE participants were selected from among the European-ancestry individuals in the UK Biobank[11], based on being in the "middles and extremes of the forced expiratory volume in 1 s (FEV_1) distribution among heavy smokers (mean 35 pack-years) and never smokers."[11] Because the UKB Axiom cohort is the complement of the UK BiLEVE cohort, the UK Axiom cohort is also non-random, being under-sampled on heavy smokers and never smokers. It is thus only the complete UKB that is a population-based study without any particular sampling scheme based on lung function and smoking, while the UK BiLEVE and UKB Axiom cohorts are selected subsamples of the UKB.

We decided to analyze the UKB as a single cohort, rather than treating the UK BiLEVE and the UKB Axiom cohorts as two separate cohorts to be analyzed separately and then included in the meta-analysis as separate cohorts. (As indicated in **Supplementary Note section 2.3**, we included fixed effects to control for the genotyping arrays in the GWAS analyses.) Several factors led us to analyze the UKB as a single cohort. First, this allowed us to control for cryptic relatedness across the BiLEVE and Axiom samples with linear mixed models (LMM) with the BOLT-LMM v2.2 software[7]. Analyzing the two cohorts separately would have necessitated dropping individuals who have relatives in the other cohort. Second, to our knowledge no published studies have analyzed the two cohorts separately. The pre-print of the UKB flagship paper[4], as well as many other recent large-scale GWAS[14–16], also analyzed the UKB as a single cohort.

During the revision stage, a Referee raised the point that our GWAS results could be sensitive to our decision of analyzing the UKB as a single cohort. Though we believe it is preferable to treat the UKB as a single cohort, it would be worrying if our results were sensitive to that decision. To verify that, we repeated our discovery GWAS of general risk tolerance and our GWAS of ever smoker, this time treating the UK BiLEVE and the UKB Axiom cohorts as two separate cohorts (and meta-analyzing the results). As we report in **Supplementary Note section 3.5**, the results barely changed.

Quality control of allele-frequency differences between the UK Biobank genotyping arrays

It was communicated that the first release of UKB data contained a small set of 65 autosomal SNPs that appeared to have flipped reference alleles contingent on the array. A subset of these had unfortunately been used during the imputation procedure. We already control for genotype array and batch during GWAS analyses, but as an additional QC step beyond excluding the 65 previously reported flipped SNPs, we investigated the allele frequencies across the arrays to be sure that our results were unaffected by artifacts from the genotyping procedure. It should be noted that the participants genotyped with the UK BiLEVE array were chosen based on lung function and smoking behavior[11], but the sample is in all other respects comparable to the rest of the UK Biobank[17].

We restricted the imputed genotype data to unrelated individuals of British ancestry to ensure that allele-frequency differences across the genotyping arrays would not be caused by differences in the proportion of ancestries or be affected by dependent observations. With PLINK[10], we calculated the allele frequencies contingent on the genotyping array for both the directly genotyped and imputed SNPs. Because our quality-control protocol, described in the next section, restricts the GWAS to SNPs with MAF ≥ 0.001, we chose not to investigate SNPs with MAF < 0.001 (in the imputed genotype data) for allele-frequency differences between the UKB genotyping arrays. SNPs available on only one of the genotyping arrays and SNPs that were not available in our main reference panel were not considered in this investigation of allele-frequency differences.

For each SNP included in the investigation, three quantities measuring differences in allele frequencies were examined: the absolute value of the difference between the two arrays and the absolute value of the difference between the main reference panel and each of the arrays. We flagged a SNP as problematic if it fulfilled the following two conditions: (1) if the absolute value of the difference between the two arrays was greater than 0.25; and (2) if the absolute value of the difference between the main reference panel and at least one of the genotyping arrays was greater than 0.25. The comparison resulted in 600 flagged autosomal SNPs (including the 65 SNPs that were already reported as problematic) that were removed from the UKB summary statistics during QC in **Supplementary Note section 2.6.2.**

Description of major steps in quality-control (QC) analyses

For each cohort, we applied a stringent quality-control protocol based on the EasyQC software (version 9.2) developed by the GIANT consortium[18], as well as additional steps developed by the Social Science Genetic Association Consortium[6,19]. All issues raised during implementation of the protocol described below were resolved through iterations between the meta-analyst and the cohort analysts before any GWAS summary statistics were forwarded for meta-analysis.

Pre-QC verification and harmonization of GWAS summary statistics

All cohorts were asked to supply descriptive statistics and phenotype definitions according to the pre-specified analysis plan[1]. The completeness of these documents was assessed as the first step of the quality control, together with examination of the uploaded GWAS summary statistics. All GWAS summary statistics were harmonized to ensure that the SNP identifier was in an admissible format (i.e. either an rsID, or in a format containing the chromosome, base pair (bp), and the two alleles of the SNP), that the missing string operator was set to “NA,” and that all files had the same column delimiter.

Filters applied before EasyQC protocol

Following recommendations provided by the UK Biobank, we removed the 65 autosomal SNPs from the UKB that had been flagged as having incorrect annotation, together with the additional 535 SNPs that we flagged in section 1.5, before applying the EasyQC protocol described below.

Also, for cohorts imputed with the September or December 2013 haplotype release of the 1000 Genomes imputation reference panel, we removed 737 SNPs with incorrect strand alignment[g].

EasyQC protocol

The filters applied in the EasyQC software are explained below in chronological order of implementation. Note that the order of the filters does not influence the outcome of the cleaned GWAS summary statistics (although it affects at which specific filter a SNP is removed). The number of SNPs filtered at each step of the EasyQC protocol is reported in Panel **A** of **Supplementary Table 25**.

Step 1 in the EasyQC protocol filtered out SNPs for which either the effect-coded allele or the other allele has values different from "A," "C," "G," or "T." This step removed all structural variants such as INDELs.

Step 2 filtered out SNPs with missing values for one or more of the following variables: *P* value, an estimated effect size (beta) or its standard error, frequency of the reference allele, imputation status, and imputation accuracy (conditional on the SNP being imputed). This filter also removed SNPs with nonsense values outside of permissible ranges such as negative or infinite standard errors, nonsensical *P* values, allele frequencies greater than 1 or below 0, as well as imputation status not equal to 1 or 0.

The thresholds chosen for the filters applied in steps 3 to 5 are summarized in **Supplementary Table 26**. Step 3 filtered out SNPs with a MAF below 0.1% for the UKB and 23andMe cohorts and below 1% for all other cohorts; this effectively removed any SNPs that were monomorphic in the summary statistics. Step 4 excluded SNPs based on imputation accuracy with a threshold contingent on the cohort-specific imputation software (0.6 for MACH, 0.7 for IMPUTE, and 0.8 for PLINK). Step 5 filtered out directly genotyped SNPs with a Hardy-Weinberg equilibrium exact test *P* value below a threshold contingent on the cohort sample size. The applied thresholds were 10^{-3} if $n < 1{,}000$, 10^{-4} if $1{,}000 \leq n < 2{,}000$, and 10^{-5} if $2{,}000 \leq n < 10{,}000$.

Two additional filters were applied to ensure that only high-quality SNPs were being forwarded to the meta-analysis; step 6 removed SNP j if

$$(2) \qquad \widehat{SE}_j > 1.4\frac{\hat{\sigma}_Y}{\sqrt{2 \cdot n_j \cdot MAF_j \cdot \left(1 - MAF_j\right)}},$$

where $\hat{\sigma}_Y$ is the standard deviation of the phenotype, n_j is the sample size, SE_j is the standard error of the coefficient estimate for SNP j, and MAF_j is the minor allele frequency of SNP j. This filter eliminates SNPs whose coefficient estimates have standard errors that are more than ~40% larger than what would be expected given the sample size, the MAF of the SNP, and the standard deviation of the phenotype. The second additional filter, step 7, removes SNPs with coefficient estimates larger than what would correspond to an R^2 greater than 5%. We adapted the filter from Okbay *et al.*[6] using an approximation to the R^2: SNP j is dropped if

[g] The announcements are available on the webpage of IMPUTE2 https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_1_integrated_SHAPEIT2_16-06-14.html and https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated_SHAPEIT2_9-12-13.html

$$|\hat{\beta}_j| > \frac{\sqrt{0.05} \cdot \hat{\sigma}_Y}{\sqrt{2 \cdot MAF_j \cdot (1 - MAF_j)}}. \tag{3}$$

Step 8 filtered out duplicate SNPs (SNPs with identical NCBI build 37 (UCSC hg19) chromosome and base-pair positions). This was implemented after the chromosome and base-pair positions of the SNPs had been harmonized with the main reference panel described above.

Step 9 aligned the SNPs to the main reference panel to ensure that the effect-coded allele was the same for all SNPs across the cohorts. This step removed SNPs that were not available in the main reference panel as well as SNPs that displayed an allele mismatch when compared to the reference (e.g., a SNP with the alleles A and T in the GWAS summary statistics would be removed if the alleles according to the reference panel were A and G).

Step 10 removed SNPs that deviated from the main reference panel in terms of allele frequency. A SNP was removed if the absolute difference between its allele frequencies in the cohort's data and in the main reference panel was greater than 0.2. Step 10 was applied to all cohorts including the UKB (for the UKB, this filter was thus applied in addition to the filter described in **Supplementary Note section 2.5** and **Supplementary Note section 2.6.2**, the purpose of which was to avoid potential strand issues caused by the two different UKB genotyping arrays).

The output from the quality control was examined to see if any filters removed an unusual or unexpected number of SNPs. Some cohorts required iterations with the analysts to ensure that all possible errors were resolved. The number of SNPs filtered at each step of the final quality control iteration is reported in Panel **A** of **Supplementary Table 25** together with the estimated genomic inflation factor (λ_{GC}).

Visual inspection of diagnostic plots

Once low-quality SNPs were filtered out, the remaining SNPs were used to produce several diagnostic plots for each cohort, most of which are the standard output of the EasyQC software. Visual inspection of these plots enabled the identification of possible issues or errors in the GWAS summary statistics of each cohort; for a more thorough discussion we refer the interested reader to Winkler *et al.*[18]. For any potential issues observed in these plots, we contacted the cohort-specific analyst and ensured that the observed issues were completely resolved. The following plots were examined:

Allele Frequency Plots (AF Plots): The AF plot contrasts the *observed* allele frequencies with the *expected* allele frequencies calculated according to the main reference panel, and the plot was created before step 10 of the EasyQC protocol. If the sample closely resembles the reference panel in terms of allele frequencies, then the SNPs should align in a diagonal with positive slope. This plot enables the analyst to detect deviations in ancestry from the reference as well as issues related to the alignment of the effect-coded allele. If the wrong effect-coded allele has been specified, then the AF plot shows a diagonal with negative slope.

P-Z Plots: Inspection of this plot shows if the reported *P* values are consistent with the reported coefficient estimates and their standard errors. One common problem observable with the P-Z plot is an erroneous column header in the GWAS summary statistics, such that the wrong column is used for either the beta estimates, standard errors or *P* values in the analysis.

Q-Q Plots: Inspection of Q-Q plots enables visualization of unaccounted-for stratification in the cohorts. No cohort displayed premature lift-off in the Q-Q plot associated with unaccounted-for stratification. The genomic inflation factors λ_{GC} are displayed in Panel **A** of **Supplementary Table 25**.

SE Plots: We plotted the observed standard errors (SE_j) of the coefficient estimates versus the standard errors expected given the MAF_j and the sample size n_j of a given SNP *j,* and the standard deviation of the phenotype $\hat{\sigma}_Y$ (which is equal to 1 if the phenotype has been standardized). This enables visual inspection to identify groups of outlier SNPs with regard to the observed standard error. The expected standard error was calculated according to the following formula:

$$\text{(4)} \qquad \widehat{SE}_j \approx \frac{\hat{\sigma}_Y}{\sqrt{2 \cdot n_j \cdot MAF_j \cdot (1 - MAF_j)}}$$

All cohorts had to pass visual inspection as well as inspection of the number of excluded SNPs at each of the exclusion filters described in the previous subsection before being passed on for the meta-analysis.

Meta-analysis, adjustment of the standard errors, and test of heterogeneity of effect sizes across cohorts

Meta-analysis

Sample-size weighted meta-analysis of the cleaned cohort-level GWAS summary statistics were carried out using the METAL software[20]. We conducted four main meta-analyses: (1) we meta-analyzed the discovery GWAS combining the UKB and 23andMe cohorts; (2) we meta-analyzed the results of the 10 replication cohorts without the UKB and 23andMe discovery cohorts, to obtain our replication GWAS; (3) we meta-analyzed the results of the 10 replication cohorts together with those of the UKB and 23andMe discovery cohorts for the follow-up analyses that use GWAS summary statistics; and (4) we meta-analyzed the results from our UKB GWAS of ever smoker with those from the TAG Consortium[2]. No meta-analyses were conducted for the five other supplementary GWAS, because data for these GWAS each came from only one cohort (either the UKB or the 23andMe cohort).

All meta-analyses were performed with the unique ID format as the identifier of each SNP (e.g., 1:123456:C:T)[19]. All meta-analyses were restricted to SNPs with a sample size greater than half of the maximum sample size across all the SNPs in the GWAS. Thus, because the discovery GWAS of general risk tolerance consists only of the 23andMe and UKB cohorts and because the 23andMe cohort is slightly larger than the UKB, all 9,284,738 SNPs available in the 23andMe cohort (and no other SNPs) were analyzed in the discovery GWAS of general risk tolerance. Of these 9,284,738 SNPs, 8,989,321 are available in both the UKB and the 23andMe cohorts and have a sample size of 931,651 or 939,908[h], and 295,417 are available only in the 23andMe cohort and have a sample size of 500,525 or 508,782[i]. Of the 124 general-risk-tolerance lead SNPs we report below in **Supplementary Note section 3.3**, all but one (rs13251864) are present in both the 23andMe and UKB cohorts. For the replication GWAS of general risk tolerance, 6,986,015 SNPs were analyzed; 9,339,358 SNPs were analyzed in the GWAS of adventurousness; and ~11,515,000 SNPS were analyzed in the GWAS of the four

[h] 8,949,622 SNPs have a sample size of 939,908 and 39,699 SNPs have a sample size of 931,651.

[i] 290,259 SNPs have a sample size of 508,782 and 5,158 SNPs have a sample size of 500,525.

risky behaviors and their first PC. Panel **B** of **Supplementary Table 25** reports the number of SNPs for all the main GWAS.

Adjustment of the standard errors

Instead of applying genomic control with the over-conservative λ_{GC}, we inflated the standard errors by the square root of the estimated intercept from an LD Score regression. This procedure allows us to correct only for inflation of test statistics caused by population stratification and other confounding factors rather than polygenicity[21].

For the discovery and replication GWAS of general risk tolerance and for the GWAS of ever smoker—all of which involved meta-analyses of cohort-level data—we only inflated the meta-level standard errors (i.e., we did not inflate the cohort-level standard errors before the meta-analysis). Likewise, for the meta-analysis of the discovery and replication GWAS for the follow-up analyses, we only inflated the meta-level standard errors. We also inflated the standard errors of the other supplementary GWAS.

In practice, for a given meta-analysis, the METAL software[20] outputs the SNPs' meta-analyzed *z*-statistics, deflated by the square root of the estimated intercept from an LD Score regression. We use SNP j's GWAS sample size n_j and minor allele frequency MAF_j, as well as the phenotype's standard deviation $\hat{\sigma}_y$, to approximate the inflated standard error of our estimate of SNP j's effect size:

$$\widehat{SE}_j \approx \sqrt{Intercept} \cdot \frac{\hat{\sigma}_y}{\sqrt{2 \cdot n_j \cdot MAF_j(1 - MAF_j)}},$$

where $\sqrt{Intercept}$ is the square-root of the LD Score intercept used to deflate the *z*-statistic in the meta-analysis.

We then used SNP j's deflated *z*-statistics $\hat{z}_j$ to approximate SNP j's effect size as $\hat{\beta}_j \approx \hat{z}_j \cdot \widehat{SE}_j$ [j]. Since the general-risk-tolerance phenotype is not measured in natural units, and since its standard deviation differs across cohorts, we normalize it to have a standard deviation of one when we estimate the SNPs' effect sizes and standard errors. For consistency, we make the same assumption when approximating the SNPs' effect sizes and standard errors for the risky behaviors and their first PC. Hence, our estimated effect sizes (the $\hat{\beta}_j$'s) are expressed in standard-deviation units of the phenotype per effect-coded allele; we hereafter refer to this as a "standardized beta," although it is only standardized in terms of standard deviation units of the phenotype (and not with respect to the genotype). The standard deviations of the phenotypes, as originally measured, are reported in **Supplementary Table 4**.

The coefficient of determination of SNP j is approximated as[22]:

[j] Since $\hat{z}_j$ is deflated and $\widehat{SE}$ is inflated by the square root of the intercept from the LD score regression, $\hat{\beta}_j$ is neither deflated nor inflated.

(5)

$$R_j^2 \approx \frac{2 \cdot MAF_j(1 - MAF_j) \cdot \hat{\beta}_j^2}{\hat{\sigma}_y^2}.$$

Evaluation of effect-size heterogeneity across cohorts for general risk tolerance

Following a Referee's suggestion, we computed Cochran's *Q* statistic for the lead SNPs of our discovery GWAS of general risk tolerance, to evaluate the heterogeneity of our estimates across the 23andMe and UKB cohorts. (We note, however, that the power of Cochran's *Q* test is limited in our setting[23,24], because the discovery meta-analysis consists of only two cohort.) In addition to examining the *P* values of Cochran's *Q* test for each lead SNP (after Bonferroni correction for the number of lead SNPs), we generated an omnibus test statistic for heterogeneity by summing the Cochran *Q* statistics across all lead SNPs[25]. Because there are two cohorts, the *Q* statistic for each lead SNP has a χ^2 distribution with one degree of freedom. The sum of these *Q* statistics is therefore (approximately) χ^2-distributed with the number of degrees of freedom being equal to the number of lead SNPs. We report the results in **Supplementary Note section 3.3.2**.

Approximately independent lead SNPs, loci, and conditional analysis

Approximately independent lead SNPs

To identify approximately independent genome-wide significant "lead SNPs", we used PLINK[10] to apply a "clumping algorithm" to the GWAS results. (We define a SNP as "genome-wide significant" if its GWAS *P* value is less than 5×10^{-8}.) Our clumping algorithm uses four parameters: a primary *P* value threshold (5×10^{-8}), a secondary *P* value threshold (1×10^{-4}), an r^2 threshold (0.1), and a SNP window defined in kilobases (1,000,000 kb). First, the SNP with the lowest *P* value (less than the primary *P* value threshold) is taken as the "lead SNP" in the first clump, and the first clump is formed by all SNPs with a *P* value smaller than the secondary *P* value threshold[k], with an r^2 greater than 0.1 with the clump's lead SNP, and within a distance less than the SNP window from the lead SNP. (We used a very wide SNP window of 1,000,000 kb, which effectively makes the r^2 and *P* value thresholds the only binding parameters for the PLINK clumping algorithm.) Next, the SNP with the second lowest *P* value (less than the primary *P* value threshold) outside the first clump becomes the lead SNP of the second clump, and the second clump is created analogously but using only the SNPs outside of the first clump. This process continues until every SNP with a *P* value less than the primary *P* value threshold is either defined as the lead SNP of a clump or clumped with another lead SNP. The r^2 was calculated with the main reference panel. Thus, a "lead SNP" is the most significant genome-wide significant SNP in an approximately independent clump, and a lead SNP cannot be in the clump of another lead SNP for the same phenotype.

Definition of non-overlapping, continuous genomic loci

For the purpose of defining non-overlapping, continuous genomic loci, we followed Ripke *et al.*[26]. Ripke *et al.* defined a locus as "the physical region containing all SNPs correlated at $r^2 >$ 0.6 with [one of the lead] SNPs", and merged associated loci within 250kb of each other into a single locus. For each of the seven GWAS, we followed this definition and created a set of loci. We report the identified loci in **Supplementary Note section 3**, and the loci are listed

[k] The secondary *P* value threshold lowers the computational effort by allowing the algorithm to ignore SNPs with large *P* values.

Supplementary Table 3 and **3.2**. In that section, we also report the loci we obtained after pooling all the loci from across the seven GWAS and merging loci within 250kb of each other; those loci are listed in **Supplementary Table 7**.

Conditional and joint multiple-SNP (COJO) analysis with GCTA

Because we consider approximately independent (pairwise $r^2 < 0.1$) lead SNPs and loci (rather than fully independent lead SNPs and loci), some of our lead SNPs could in principle be secondary associations that are driven by their LD with extremely strong primary associations. We thus performed conditional and joint multiple-SNP (COJO) analysis[27] with GCTA. For each of the seven GWAS, we restricted the analysis to the set of SNPs that (1) pass all GWAS quality control filters, and (2) are located within the loci of the phenotype (which includes all the lead SNPs). We analyzed the summary statistics using the stepwise model-selection algorithm detailed in the original COJO publication[27]. The analysis requires two input parameters: (1) the distance in kb at which perfect linkage equilibrium ($r^2 = 0$) is assumed, and (2) an r^2 threshold that prevents the stepwise model selection from adding SNPs that are highly correlated with a previously selected SNP. We used the default parameters, which assume perfect linkage equilibrium for SNPs separated by 10 Mb and which do not add SNPs in strong LD ($r^2 > 0.9$) with an already selected SNP. The COJO analysis was performed with LD estimated in our main reference panel (described in **Supplementary Note section 2.4**). We report the results of the COJO analysis in **Supplementary Note section 3**.

As we discuss in **Supplementary Note section 3.6**, we also conducted a multiple regression analysis with individual-level data from the UKB. In that analysis, for each phenotype (except adventurousness, for which there is no UKB data), for each chromosome we regressed the phenotype on all the phenotype's lead SNPs located on the chromosome (and on control variables). The results were consistent with those of the COJO conditional analysis.

Check for long-range LD regions, candidate inversions, and 1000 Genomes structural variants

Check for long-range LD regions

We investigated if the lead SNPs were located in larger structural variation in the form of long-range LD regions, by using a set of long-range LD regions from Price *et al.*[28]. Price *et al.* identified 24 long-range LD regions from 327 European-ancestry individuals, which replicated in two independent samples comprising 1593 European-ancestry Americans and 3004 British individuals. We lifted the genomic positions of the long-range LD regions to build 37 (GRCh37) with the UCSC genome-annotation lift-over tool[1], and there were nine long-range LD regions that could not be lifted due to non-overlapping genome sequences or ambiguous mapping across the builds. Hence, the combined map contains 15 long-range LD regions that have non-ambiguous genomic positions available in build 37. These range from ~2.5 to ~8 Mb in genomic size.

We checked if each lead SNP from the GWAS was within a long-range LD region, or within 250 bp from the breakpoints of such a region. The results are reported in **Supplementary Table 3** and **Supplementary Table 6**.

[1] The lift-over tool is available here: https://genome.ucsc.edu/cgi-bin/hgLiftOver

Check for candidate inversions

We also investigated if the lead SNPs were located in larger structural variation in the form of candidate inversions, by using a list of genomic segments highly prone to inversion polymorphisms from an unpublished resource by Gonzalez, J.R. & Esko, T., (2017, unpublished). The genomic segments highly prone to inversion polymorphisms were identified based on the knowledge that submicroscopic human inversions are typically flanked by highly homologous flanking repeats[29], which predisposes their occurrence by non-allelic homologous recombination. Therefore, a set of segments was selected that may be prone to submicroscopic inversions, consisting of all single copy segments in the Genome Reference Consortium's human reference sequence build 36 (GRCh36) between 0.1 and 8 Mb in length, and flanked by segmental duplications with 90% identity (across the flanking duplications). In total, there were 173 segments that met these criteria and that were thus considered as genomic segments highly prone to inversions. As detectable traces of inversions in SNPs depend on many factors—such as being frequent, ancient and nonrecurring—we tested whether the segments showed inversion patterns in any of two different SNP datasets. First, 69 (40%) of the segments overlapped with the inversions that Caceres *et al.*[30] obtained in the phased genotypes of CEU individuals from the HapMap III project. Second, inversion-like haplotypes[31] were inferred in a subsample of 882 Estonians for which gene expression data was available in peripheral blood. In this case 65 (38%) of the 173 segments were significantly associated with the expression of single copy genes within the segment. In total 104 (60%) of the 173 segments showed an inversion signal, indicating their predisposition for inversion occurrence.

We lifted the genomic positions of the genomic segments highly prone to inversion polymorphisms to build 37 (GRCh37) with the UCSC genome-annotation lift-over tool[m], and there were 19 genomic segments that could not be lifted due to non-overlapping genome sequences or ambiguous mapping across the builds. Hence, the combined map contains 154 segments prone to inversion polymorphisms (hereafter referred to as "the 154 candidate inversions"), that have non-ambiguous genomic positions available in build 37. These range from ~500 kb to ~8 Mb in genomic size.

We checked if each lead SNP from the GWAS was within such a candidate inversion, or within 250 bp from the breakpoints of such a candidate inversion. The results are reported in **Supplementary Table 3** and **Supplementary Table 6**.

Check for 1000 Genomes structural variants

Sudmant *et al*. (2015)[32] called and classified a large number of structural variants (SVs) with the final version of the 1000 Genomes Project phase 3 reference panel. They have released an integrated map of 37,250 smaller structural variants, together with enhanced resolution of the size and breakpoint compared to previous publications, and we hereafter refer to these as the "1000G structural variants." The structural variants range from 1 bp to ~445 kb in genomic size. The smallest variants are generally insertions of 1 bp (~6,900 variants), and the larger variants are generally deletions larger than 50 bp. The majority of these structural variants are in LD with proximate SNPs, and we therefore checked if any of the lead SNPs from our main GWAS, and SNPs in strong LD with those lead SNPs, were located within the start and end positions of any of the 37,250 structural variants. We defined strong LD as an r^2 greater than 0.8, which is the definition used by the 1000 Genomes Project Consortium[32]. The SNPs in LD were extracted using PLINK[10] and the main reference panel. We report the results in **Supplementary Table 3** and **Supplementary Table 6**.

[m] The lift-over tool is available here: https://genome.ucsc.edu/cgi-bin/hgLiftOver

Investigation of the novelty of our GWAS associations

To investigate the novelty of our GWAS associations, we performed lookups of our lead SNPs (and the SNPs in LD with the lead SNPs, $r^2 > 0.1$) in the NHGRI-EBI GWAS Catalog database (revision 2017-08-15)[33] of genome-wide significant associations from previous GWAS. We also looked up our lead SNPs (and the SNPs in LD, $r^2 > 0.1$) in some recent GWAS articles that have not been catalogued in the NHGRI-EBI GWAS Catalog database. The NHGRI-EBI GWAS Catalog is a resource that aims to catalogue all associations reported in published GWAS.

For general risk tolerance, we performed a search with the term "risk" in the index of phenotypes, and we did not find any previous studies on general risk tolerance in the Catalog. We know of one previous study that identified one independent genome-wide significant association with general risk tolerance, and of one concurrent study that identified a second genome-wide significant association, both using the first UKB data release[34,35]; the authors of ref.[34] referred to the phenotype as "risk-taking propensity," and the authors of ref.[35] referred to it as "risk-taking behavior." The first genome-wide significant association is replicated in an online publication published in advance[36]. We added the first of these two studies (i.e., Day *et al.*[34]) to our investigation of the novelty of our general-risk-tolerance lead SNPs, and the second we consider concurrent. We also note that, in **Supplementary Note section 11.1**, we report the results of a literature search of association studies of risk tolerance; that literature search identified no previously reported genome-wide significant associations.

The phenotypes drinks per week, ever smoker, and number of sexual partners (or related phenotypes such as alcoholism and age at first sex), were available in the NHGRI-EBI GWAS Catalog database (revision 2017-08-15). Since the GWAS Catalog is not always up-to-date, we additionally performed a literature search for genome-wide significant findings that might not yet have been included in the Catalog. We searched the Pubmed literature database on March 6 2017 and September 13 2017, for the term "genome-wide association study" together with each of the terms "alcohol," "sexual," and "smoking" individually. We screened the abstracts and compared the resulting articles with the article list of the GWAS catalog. In addition to what was already reported in the GWAS Catalog (revision 2017-08-15), we found three additional studies with genome-wide significant findings on alcohol consumption, two additional studies on smoking, and no additional studies on sexual behaviors.

To our knowledge, this is the first GWAS of adventurousness, automobile speeding propensity, and of the first PC of the four risky behaviors; unsurprisingly, we could not find previous GWAS on any of these phenotypes in the NHGRI-EBI GWAS Catalog database (revision 2017-08-15).

GWAS catalog lookup

We investigated whether the lead SNPs from our discovery GWAS of general risk tolerance and from our supplementary GWAS have previously been associated at genome-wide significance with any phenotypes in the NHGRI-EBI GWAS Catalog database[33] (revision 2017-08-15). We query the GWAS Catalog with our list of lead SNPs, and the SNPs in LD with a lead SNP ($r^2 > 0.6$).

Since the NGRI-EBI GWAS Catalog is not always up-to-date with results from the most recent publications (especially in non-peer reviewed outlets such as BioRxiv), we also queried the summary statistics of the most recently published GWAS on attention deficit hyperactivity disorder[37], autism spectrum disorder[38], and anorexia nervosa[n]. We also queried the genome-wide significant findings from the additional GWAS on alcohol intake and smoking, which we found in addition to the GWAS Catalog, as detailed in the previous section. We queried the additional GWAS on alcohol intake and smoking because they contained at least one genome-wide significant result; because their results were publicly available; and because the meta-analysis that produced them did not include the UK Biobank (that comprise our discovery sample together with 23andMe).

We perform these lookups because the existence of SNPs and genes associated with both one of our studied phenotypes and another phenotype can point to a common genetic etiology. However, it is important to note that two phenotypes that share a genetic locus do not necessarily have to share the same causal *variant* at that locus due to the widespread LD that characterizes the human genome. Moreover, even if two phenotypes do share a single causal variant, they do not have to share general underlying genetic etiologies. For instance, recent work[39] has shown that two types of autoimmune diseases (rheumatoid arthritis and ulcerative colitis/Crohn's disease) that are known to share risk loci are not genetically correlated at a genome-wide level. Here, the reason was the lack of an overall directional trend: some risk alleles for one disease were also risk alleles for the other disease, but some alleles that were *protective* for the one disease were risk alleles for the other. This resulted in a near-zero correlation at the genome-wide level. Thus, we emphasize that the current lookup does not make it possible to determine *etiological* overlap, but only hints at overlapping loci, between general risk tolerance or the supplementary GWAS phenotypes, and the phenotypes reported in the GWAS Catalog.

Cross-lookup of GWAS results

We performed cross-lookups of the lead SNPs across our discovery GWAS of our primary and supplementary phenotypes. Specifically, for each lead SNP in each of the GWAS, we checked if the SNP is in LD (with an r^2 greater than 0.1) with lead SNPs in the other GWAS. LD was calculated using PLINK[10] and the main reference panel. The results of the cross-lookups are reported in **Supplementary Table 3** and **Supplementary Table 6**. As we describe above, we also investigated if there were any long-range LD regions or candidate inversions that contained lead SNPs for multiple GWAS.

Gene annotation

We annotated the lead SNPs with gene information using the National Institute of Health (NIH) National Center for Biotechnology Information (NCBI) gene ontology database (version 2016-05-25)[o]. As the general rule, a SNP was annotated to its most proximate gene. If a SNP was located between two genes, then we compared the distance to the end coordinate of the gene upstream with the distance to the start coordinate of the gene downstream to find the most proximate gene. If a SNP was located within multiple overlapping genes, then the SNP was annotated to the gene with the most proximate start coordinate. This means that all lead SNPs were annotated to a single gene. The approach roughly partitions the SNPs throughout the genome into separate genomic segments. The annotations are displayed in **Supplementary Table 3** and **Supplementary Table 6**, where we also indicate if a lead SNP is located within

[n] The Psychiatric Genomics Consortium's GWAS summary statistics for attention deficit hyperactivity disorder, autism spectrum disorder, and anorexia nervosa (referred to as "ED," i.e. eating disorder) are publicly available and can be downloaded here: https://www.med.unc.edu/pgc/results-and-downloads.

[o] The NCBI gene ontology database is available here: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/.

or outside the gene's start and end coordinates. We also checked if there were genes to which lead SNPs from multiple GWAS were annotated (the results are reported in **Supplementary Note section 3.2**).

Methods to estimate genome-wide SNP heritability

Risk tolerance (both self-reported and experimentally elicited) has been found to be moderately heritable in twin studies, with heritability estimates ranging from 20% to 60%[40–42]. In this section, we employ three different methods to obtain estimates of the SNP heritability of our primary and supplementary GWAS phenotypes. A phenotype's SNP heritability is the fraction of the phenotype's variance that is accounted for by the additive genetic effects of a set of SNPs.

We used three methods, GCTA[43], LD Score regression[21], and Heritability Estimator from Summary Statistics (HESS)[44], to estimate the genome-wide SNP heritability (h_G^2). The GCTA method estimates the heritability of a phenotype directly from the individuals' genotypic data, while the LD Score and HESS methods use GWAS summary statistics as inputs.

For comparability across phenotypes and methods, for the LD Score and HESS methods, we used summary statistics from the UKB GWAS only for all phenotypes except adventurousness. For the adventurousness phenotype, we only report estimates that were obtained using the LD Score regression and HESS methods using the 23andMe summary statistics, because this phenotype is not available in the UKB and we did not have access to the individual-level genotypic data from the 23andMe cohort (and so could not obtain estimates with the GCTA method).

We also computed HESS SNP heritability estimates using summary statistics from our seven main GWAS (and not only from GWAS conducted in the UKB or in the 23andMe cohort).

The GCTA method

The GCTA method is based on restricted maximum-likelihood estimation and uses the genetic relationship matrix (GRM) to estimate the SNP heritability. Under the assumptions discussed in Yang *et al.* (2011)[43], the method leads to unbiased estimates of the genome-wide SNP heritability. However, it is computationally intensive, and it is thus necessary to limit the number of SNPs and individuals included in the analysis in order to be computationally feasible. Therefore, we restricted the GCTA analysis to a random subset of 30,000 individuals out of the full sample from the discovery GWAS. We thereafter dropped one individual in each pair of individuals with a cryptic relatedness exceeding 0.025, to obtain a set of unrelated individuals. For comparability we used the same initial subset of 30,000 individuals for the GCTA estimation for all phenotypes, though the sample size varies slightly across phenotypes because of missing phenotypic observations. The final sample sizes for each phenotype are presented in **Supplementary Table 30**. In total 646,855 directly genotyped SNPs with MAF > 0.01 were included in the GCTA heritability estimation.

The LD Score regression method

Under the assumptions discussed in Bulik-Sullivan *et al.* (2015)[21], a SNP's GWAS χ^2 statistic is linearly related to its LD score, defined as the sum of the squared correlation coefficients between any single SNP and all the other SNPs. The slope of the LD Score regression (of the SNPs GWAS χ^2 statistics on their LD scores and an intercept) can be rescaled to obtain an estimate of the heritability explained by the SNPs included in the LD Score analysis by dividing the slope by the sample size divided by the number of SNPs, i.e., by *n*/M. We used the "eur_w_ld_chr/" files of LD scores computed by Finucane *et al.*[45] and made available on https://github.com/bulik/ldsc/wiki/Genetic-Correlation, accessed on March 14, 2016. These LD scores were computed with genotypes from the European-ancestry samples in the 1000 Genomes Project. Only HapMap3 SNPs with MAF > 0.01 were included in the LD Score regression; for every phenotype, ~1.3 million SNPs were used for the LD Score heritability estimation. Since Genomic Control (GC) will tend to bias the intercept of the LD Score regression downward, we did not apply GC to the summary statistics prior to estimating the LD Score regressions.

The HESS method

The HESS estimator can be described in brief as an analytical variance decomposition method that, unlike the GCTA and LD Score regression methods, assumes that the SNP effect sizes are fixed effects rather than random effects. The method assumes that the SNPs are randomly distributed in the population and requires a pre-specified SNP covariance matrix as input. The SNP covariance matrix can be estimated in the sample of interest if individual genotypic data is available, or with an external reference panel such as the 1000 Genomes. As Shi *et al.*[44] show using simulations, heritability estimates from LD Score regression are sensitive to the true proportion of causal SNPs, and the HESS estimator yields more accurate heritability estimates than LD Score regression under a wider range of proportions of truly causal SNPs. We used the reference panel distributed with the HESS software for the calculation of the covariance matrix. That panel is the European subsample of the 1000 Genomes phase 3 version 5 reference panel, restricted to common variants (MAF > 0.05), which is the same as the reference panel used for the construction of the LD Scores[p]. For every phenotype, a total of ~4.9 million SNPs were used in the HESS heritability estimation. As with the LD Score regressions, we did not apply GC prior to estimating heritability with HESS.

[p] While the same reference panel was used for the construction of the LD scores, as indicated above HapMap3 SNPs with MAF > 0.01 were included in the LD score regression.

References – Chapter 2 Supplementary methods

1. SSGAC. Pre-registered Analysis Plan - GWAS Risk tolerance. *Open Science Framework* (2016). Available at: https://osf.io/cjx9m/.
2. The Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42,** 441–7 (2010).
3. Marchini, J. L. *et al. Genotype imputation and genetic association studies of UK Biobank*. (2015).
4. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017). doi:http://dx.doi.org/10.1101/166298
5. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).
6. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).
7. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).
8. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31,** 2202–2204 (2015).
9. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6,** 8111 (2015).
10. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
11. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank. *Lancet Respir. Med.* **3,** 769–781 (2015).
12. 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 (2015).
13. Marchini, J. L. *et al. Genotype Imputation and Genetic Association Studies of Uk Biobank: Interim Data Release*. (2015).
14. Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany. NY).* **9,** 2504–2520 (2017).
15. Jansen, P. R. *et al.* Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. *bioRxi* (2018).
16. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *bioRxiv* (2018). doi:10.1101/274654
17. UK Biobank. *Genotyping and Quality Control of UK Biobank, a Large-Scale, Extensively Phenotyped Prospective Resource: Information for Researchers. Interim Data Release.* (2015).
18. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9,** 1192–212 (2014).

19. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624–633 (2016).

20. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).

21. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

22. Rietveld, C. A. C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science (80-.).* **340,** 1467–1471 (2013).

23. Pereira, T. V., Patsopoulos, N. A., Salanti, G. & Ioannidis, J. P. A. Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. *Res. Synth. Methods* **1,** 149–161 (2010).

24. Ioannidis, J. P. A., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2,** (2007).

25. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10,** 101–129 (1954).

26. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

27. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–375 (2012).

28. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83,** 132–139 (2008).

29. Cáceres, A. & González, J. R. Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res.* **43,** e53 (2015).

30. Cáceres, A., Sindi, S. S., Raphael, B. J., Cáceres, M. & González, J. R. Identification of polymorphic inversions from genotypes. *BMC Bioinformatics* **13,** 28 (2012).

31. Ma, J. & Amos, C. I. Investigation of inversion polymorphisms in the human genome using principal components analysis. *PLoS One* **7,** e40224 (2012).

32. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

33. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-1006 (2014).

34. Day, F. R. *et al.* Physical and neurobehavioral determinants of reproductive onset and success. *Nat. Genet.* **48,** 617–623 (2016).

35. Strawbridge, R. J. *et al.* Genome-wide analysis of self-reported risk-taking behaviour and cross-disorder genetic correlations in the UK Biobank cohort. *Transl. Psychiatry* **8,** 1–11 (2018).

36. Boutwell, B. *et al.* Replication and characterization of CADM2 and MSRA genes on human behavior. *Heliyon* **3,** e00349 (2017).

37. Demontis, D. *et al.* Discovery of the first genome-wide significant risk loci for ADHD. *bioRxiv* (2017). doi:10.1101/145581

38. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8,** 21 (2017).

39. Bulik-Sullivan, B. K. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).

40. Beauchamp, J. P., Cesarini, D. & Johannesson, M. The psychometric and empirical properties of measures of risk preferences. *J. Risk Uncertain.* **54,** 203–237 (2017).

41. Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P. & Wallace, B. Genetic variation in preferences for giving and risk taking. *Q. J. Econ.* **124,** 809–842 (2009).

42. Zyphur, M. J., Narayanan, J., Arvey, R. D. & Alexander, G. J. The genetics of economic risk preferences. *J. Behav. Decis. Mak.* **22,** 367–377 (2009).

43. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

44. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99,** 139–153 (2016).

45. Finucane, H. K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

46. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).

47. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).