**GWAS-based pathway analysis differentiates between fluid and crystallized intelligence**

Andrea Christoforou, Thomas Espeseth, Gail Davies, Carla P.D. Fernandes, Sudheer Giddaluru, Manuel Mattheisen, Albert Tenesa, Sarah E. Harris, David C. Liewald, Antony Payton, William Ollier, Michael Horan, Neil Pendleton, Paul Haggarty Srdjan Djurovic, Stefan Herms, Per Hoffman, Sven Cichon, John M. Starr, Astri Lundervold, Ivar Reinvang, Vidar M. Steen, Ian J. Deary, Stephanie Le Hellard

**Supporting Information**

**p. 2** **Supporting Information Methods**

**p. 4** **Supporting Information Figure S1**

**p. 5** **Supporting Information Figure S2**

**p. 6** **Legends for Supporting Information Figures**

**pp. 7-15** **Descriptions of Supporting Information Datasets**

**p. 16** **Supporting Information References**

**Supporting Information Methods**

*Cognitive phenotypes: Tests used to construct* g*F and* g*C.*

As described in Davies et al. (2011), cognitive phenotypes of crystallised (*g*C) and fluid (*g*F) intelligence were constructed for each of the five CAGES cohorts and the NCNG sample. The following tests were used to represent *g*C: The National Adult Reading Test in the Lothian Birth Cohorts of 1921 and 1936, and the Aberdeen Birth Cohort 1936; the Mill Hill Vocabulary Test in the Manchester and Newcastle samples; and the Wechsler Adult Intelligence Scale Vocabulary subtest in the NCNG sample. For *g*F, principal components analyses (PCA) was applied to the scores on various *gF*-type cognitive tests to derive a single *g*F factor in the following cohorts: the Lothian Birth Cohorts 1921 and 1936, the Aberdeen Birth Cohort 1936, and the NCNG sample. The tests used to form the *g*F factor in the LBC1921 were the Moray House Test, Raven's Matrices, Logical Memory, and Verbal Fluency. For the LBC1936, the six tests from the WAIS-III$^{UK}$ were used. The ABC1936 *g*F factor included Raven's Progressive Matrices, Digit Symbol, Uses of Common Objects, and AVLT. Individuals' scores on the first unrotated principal component were extracted and used as the indicator of *g*F. For the NCNG sample, a hierarchy of PCA analyses was used. The unrotated first component for three subtests of the California Verbal Learning Test II (Delis et al., 2000) (learning and memory) defined a memory factor. The first component from the four conditions of D-KEFS Color Word interference (Stroop test) (Trenerry 1989) defined a speed factor. These factor scores, together with the raw score of the Matrix Reasoning subscale of the Wechsler Abbreviated Scale of Intelligence (Wechsler 1999) and the overall mean of median reaction times from a multiple choice reaction time task (Espeseth et al., 2006), were used as input for a further PCA, of which the un-rotated first component defined the *g*F factor. In the Manchester and Newcastle samples, empirical Bayes's estimates for each individual

were obtained from a random effects model fitted by maximum likelihood (ML) to the standardized age-regressed residuals obtained for each sex from the Alice Heim 4 test and the Cattell Culture Fair test scores.

# Figure S1

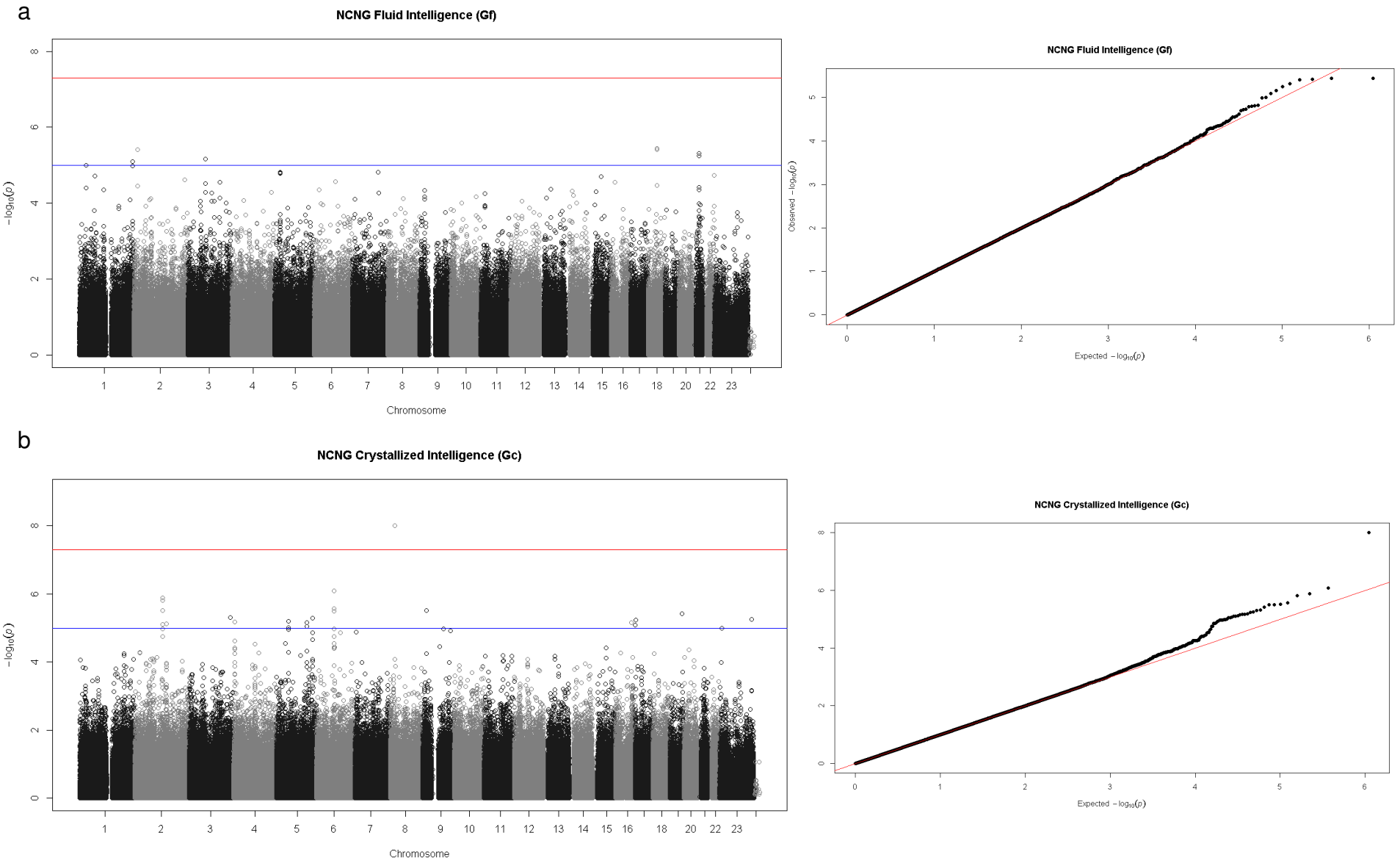
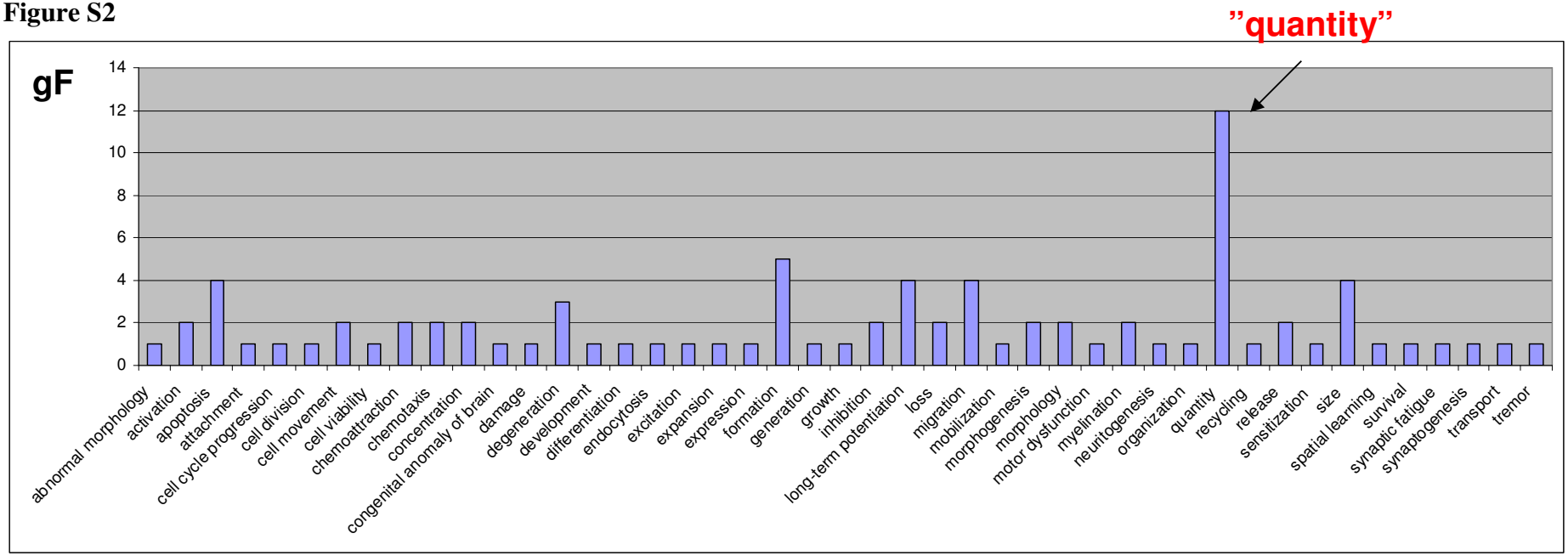
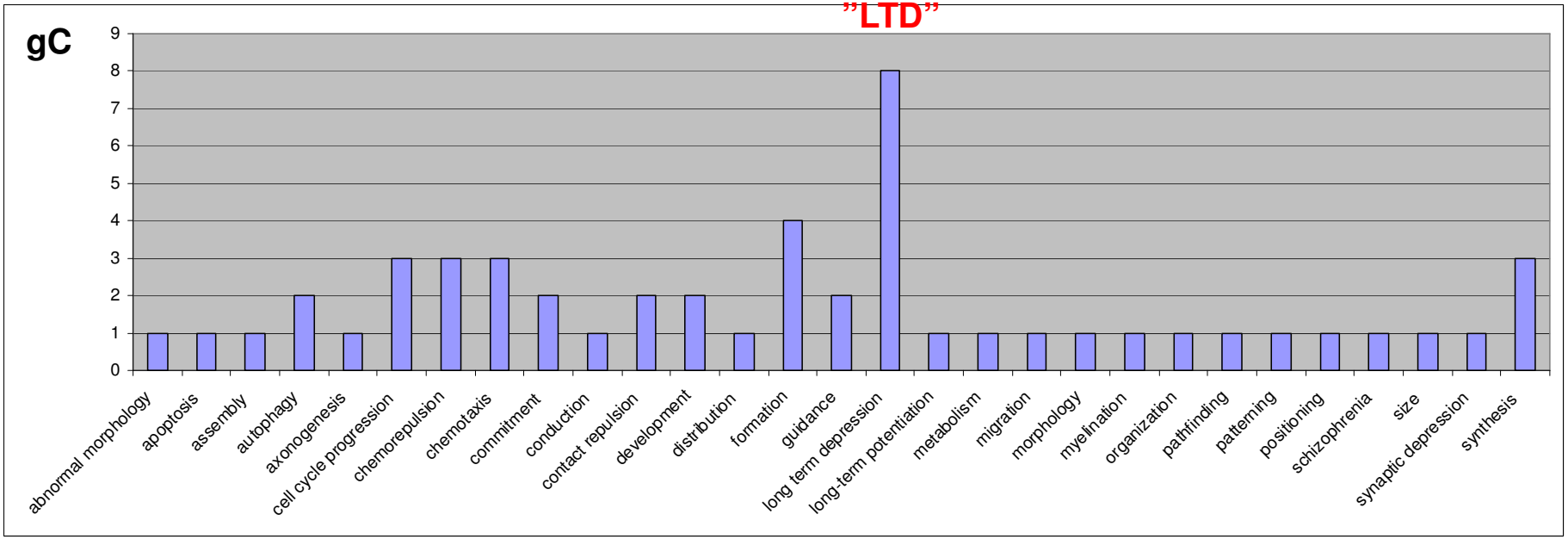

a
NCNG Fluid Intelligence (Gf)
$-\log_{10}(p)$
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23
NCNG Fluid Intelligence (Gf)
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
b
NCNG Crystallized Intelligence (Gc)
$-\log_{10}(p)$
Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23
NCNG Crystallized Intelligence (Gc)
Expected $-\log_{10}(p)$

**Figure S2**


"quantity"
gF
14
12
10
8
6
4
2
0
abnormal morphology
activation
apoptosis
attachment
cell cycle progression
cell division
cell movement
cell viability
chemoattraction
chemotaxis
concentration
congenital anomaly of brain
damage
degeneration
development
differentiation
endocytosis
excitation
expansion
expression
formation
generation
growth
inhibition
long-term potentiation
loss
migration
mobilization
morphogenesis
morphology
motor dysfunction
myelination
neuritogenesis
organization
quantity
recycling
release
sensitization
size
spatial learning
survival
synaptic fatigue
synaptogenesis
transport
tremor


"LTD"
gC
9
8
7
6
5
4
3
2
1
0
abnormal morphology
apoptosis
assembly
autophagy
axonogenesis
cell cycle progression
chemorepulsion
chemotaxis
commitment
conduction
contact repulsion
development
distribution
formation
guidance
long term depression
long-term potentiation
metabolism
migration
morphology
myelination
organization
pathfinding
patterning
positioning
schizophrenia
size
synaptic depression
synthesis

**Legends for Supporting Information Figures**

**Figure S1. Manhattan (left) and quantile-quantile (QQ; right) plots of (a) fluid and (b) crystallized intelligence in the NCNG.** The genomic inflation factor, based on the median chi-squared, as implemented in PLINK, is 1 for both fluid and crystallized intelligence. The blue line in the Manhattan plots indicates the suggestive $P=1x10^{-5}$ threshold, while the red line indicates the traditional genome-wide $5x10^{-8}$ threshold.

**Figure S2. Histogram of number of function annotations relating to each general function for *g*F (top) and *g*C (bottom).** The predominant function for *g*F ("quantity") and *g*C ("LTD") are indicated.

**Descriptions of Supporting Information Datasets (.xls files)**

**Dataset 1. Results of single-marker analyses.**

This Dataset show the top 1000 SNPs in NCNG for *g*F and *g*C, and the SNPs that meet replication criteria together with the corresponding genes for IPA analysis.

*Contains four datasheets:*

*gF Top 1000 SNPs in NCNG (Sheet S1)*

*gC Top 1000 SNPs in NCNG (Sheet S2)*

*gF 816 replicated SNPs (Sheet S3)*

*gC 884 replicated SNPs (Sheet S4)*

S1. *g*F Top 1000 SNPs in NCNG. For each SNP, the table shows the rs identifier, chromosome number, and position (NCBI bld 36). For the 629 NCNG individuals with gF measurements, the minor allele frequency (MAF) is listed, together with the identity of the minor/reference allele (Allele 1) and other allele (Allele 2). The association results for the NCNG are presented as beta/effect size, standard error (SE) and P-value. Highlighted/bold SNPs are those that show evidence of association (P-value $\leq 0.05$) and the same direction of effect in the NCNG and CAGES (see S3). The gene(s), if any, to which each of these SNPs was assigned, using LDsnpR, is shown in the final column. SNPs that could not be assigned to genes are indicated by "N/A".

S2. *g*C Top 1000 SNPs in NCNG. The table shows the rs identifier, chromosome number and position (NCBI bld 36) of each SNP. For the 643 NCNG individuals with gC measurements, the minor allele frequency (MAF) is presented together with the identity of the minor/reference allele (Allele 1) and the other allele (Allele 2). Association results for the

NCNG are presented as beta/effect size, standard error (SE) and P-value. Highlighted/bold SNPs are those that show evidence of association (P-value ≤ 0.05) and the same direction of effect in the NCNG and CAGES (see S4). The gene(s), if any, to which each of these SNPs was assigned, using LDsnpR, is shown in the final column. SNPs that could not be assigned to genes are indicated by "N/A".

S3. *g*F replicated SNPs. The table lists 816 SNPs that showed evidence of association in the NCNG and of replication in the CAGES for fluid intelligence (*g*F). For each SNP, the table includes the rs identifier, chromosome number and position (NCBI bld 36). For the 629 individuals with gF measurements, the minor allele frequency (MAF) is shown, as is the identity of the minor/reference allele (Allele 1) and the other allele (Allele 2). The association results for the NCNG are presented as beta/effect size, standard error (SE) and P-value. Also shown are the corresponding association results for the CAGES (Davies et al., 2011), with the minor/reference allele (Allele 1) indicated, and the combined meta-analysis of both the NCNG and the CAGES. Finally, the gene(s), if any, to which each of these SNPs was assigned, using LDsnpR, is shown in the final column. SNPs that could not be assigned to genes are indicated by "N/A". Bold indicates genes with SNPs that survived the stringent $P \leq 1x10^{-5}$ cutoff.

S4. *g*C replicated SNPs. The 884 SNPs that showed evidence of association in the NCNG and of replication in the CAGES for crystallized intelligence (*g*C) are listed. The rs identifier, chromosome and position (NCBI bld 36) of each SNP is presented. For the 643 NCNG individuals with *g*C measurements, the minor allele frequency (MAF) is displayed, together with the identity of the minor/reference allele (Allele 1) and the other allele (Allele 2). Association results for the NCNG are presented as beta/effect size, standard error (SE) and P-

value. Also shown are the corresponding association results (meta-P-values) for the CAGES (Davies et al., 2011) with the minor/reference allele (Allele 1) indicated, and the combined meta-analysis of both the NCNG and the CAGES. Finally, the gene(s), if any, to which each of these SNPs was assigned, using LDsnpR, is shown in the final column. SNPs that could not be assigned to genes are indicated by "N/A". Bold indicates genes with SNPs that survived the stringent $P \leq 1x10^{-5}$ cutoff.

**Dataset 2. Results of gene-based analyses.**

This Dataset shows genes with best permuted P≤0.05 in NCNG and genes meeting replication criteria and therefore selected for IPA analysis. The final sheet lists all genes tested.

*Contains 7 datasheets:*

*gF genes perm P≤0.05 in NCNG (Sheet S5)*

*gC genes perm P ≤0.05 in NCNG (Sheet S6)*

*gF 841 replicated genes (Sheet S7)*

*gC 920 replicated genes (Sheet S8)*

*gF 178 replicated genes str (Sheet S9)*

*gF 178 replicated genes str (Sheet S10)*

*All genes tested (Sheet S11)*

S5. Genes with best permuted P-value ≤ 0.05 in NCNG for *g*F. For each of the 2698 genes or ENSEMBL 54 gene identifiers, the table includes chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and the lowest permutation-based P-value from the three models tested.

S6. Genes with best permuted P-value ≤ 0.05 in NCNG for *g*C. For each of the 2615 genes or ENSEMBL 54 gene identifiers, the table includes chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and the lowest permutation-based P-value from the three models tested.

S7. Replicated genes for *g*F. The table lists the 841 genes with the most significant permutation-based P-values in the NCNG, which also show replication in the CAGES under the relaxed criteria of the unadjusted minP approach. Also included are chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES, the lowest permutation-based P-value from the three models tested in the NCNG and the minimum (min) P-value in the CAGES.

S8. Replicated genes for *g*C. The table lists the 920 genes with the most significant permutation-based P-values in the NCNG, which also show replication in the CAGES under the relaxed criteria of the unadjusted minP approach. Also included are chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES, the lowest permutation-based P-value from the three models tested in the NCNG and the minimum (min) P-value in the CAGES.

S9. Replicated genes for *g*F – stringent criteria. The table lists the 178 genes with the most significant permutation-based P-values in the NCNG, which also show replication in the CAGES under the more stringent replication criteria. Also included are chromosome (CHR)

and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES, the lowest permutation-based P-value from the three models tested in the NCNG and the lower of the two P-values obtained from either the modified Sidak correction or Brown's approximation method (CAGES adjP).

S10. Replicated genes for *g*C – stringent criteria. The table lists the 224 genes with the most significant permutation-based P-values in the NCNG, which also show replication in the CAGES under the more stringent replication criteria. Also included are chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES, the lowest permutation-based P-value from the three models tested in the NCNG and the lower of the two P-values obtained from either the modified Sidak correction or Brown's approximation method (CAGES adjP).

S11. All genes tested. The table lists all 34,109 Ensembl 54 genes tested. Also included are chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES.

**Dataset 3. IPA results**

This Dataset shows significant genes identified in IPA database for pathway analysis and the resulting significant function annotations emerging from various analyses. A final sheet is provided with a list of the "IPA Ready" genes available among all 34,109 genes tested (see Dataset 2, Sheet S9).

*Contains 11 datasheets*

*gF 853 "IPA Ready" genes (Sheet S12)*

*gC 893 "IPA Ready" genes (Sheet S13)*

*gF IPA CNS Full (Sheet S14)*

*gC IPA CNS Full (Sheet S15)*

*gF IPA CNS NR and 2+genes (Sheet S16)*

*gC IPA CNS NR and 2+genes (Sheet S17)*

*gF IPA CNS str 2+genes (Sheet S18)*

*gC IPA CNS str 2+genes (Sheet S19)*

*gF Full IPAannotations (Sheet S20)*

*gC Full IPAannotations (Sheet S21)*

*All 19,275 "IPA Ready" genes (Sheet S22)*

S12. gF "IPA Ready" genes. Of 1182 genes identified for *g*F and submitted to IPA for pathway analysis, 853 were in the IPA database and thus available for pathway analysis. For each of these genes, the ENSEMBL Id is shown, together with the Gene Symbol, the Entrez Gene Name and an indication of whether the gene was also identified for *g*C.

S13. gC "IPA Ready" genes. Of 1294 genes identified for *g*C and submitted to IPA for pathway analysis, 893 were in the IPA database and thus available for pathway analysis. For each of these genes, the ENSEMBL Id is shown, together with the Gene Symbol, the Entrez Gene Name and an indication of whether the gene was also identified for *g*F.

S14. The full IPA output for *g*F when restricting analysis to the CNS. The table includes the category, the general function, the specific function annotation (FA), the Fisher Exact Test

(FET) P-value, the corrected Benjamini-Hochberg (BH) P-value, the number of genes responsible for the enrichment signal and their gene names. The table also indicates whether the FA remains significant when BDNF is excluded or when the IPA is re-run with all cell lines (excluding cancer cell lines; see S15). The final column shows whether the FA is also significant in the corresponding *g*C analysis.

S15. The full IPA output for *g*C when restricting analysis to the CNS. The table includes the category, the general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the corrected Benjamini-Hochberg (BH) P-value, the number of genes responsible for the enrichment signal and their gene names. FAs that remain significant when the IPA is re-run with all cell lines (excluding cancer cell lines; see S16) are indicated. The final column shows whether the FA is also significant in the corresponding *g*F analysis.

S16. The non-redundant list of significant function annotations for *g*F identified on the basis of *at least* two genes. The table includes the general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the corrected Benjamini-Hochberg (BH) P-value, the number of genes responsible for the enrichment signal and their gene names. FAs that remain significant when BDNF is excluded or when the IPA is re-run with all cell lines (excluding cancer cell lines; see S20) are indicated. The final column shows whether the FA is also significant in the corresponding *g*C analysis.

S17. The non-redundant list of significant function annotations for *g*C identified on the basis of *at least* two genes. The table includes the general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the corrected Benjamini-Hochberg (BH) P-value, the number of genes responsible for the enrichment signal and their gene names. FAs that

remain significant when the IPA is re-run with all cell lines (excluding cancer cell lines; see S21) are indicated. The final column shows whether the FA is also significant in the corresponding *g*F analysis.

S18. The list of significant function annotations for *g*F identified on the basis of *at least* two genes following IPA analysis on the gene list constructed using more stringent criteria. The table includes the category, general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the number of genes responsible for the enrichment signal and their gene names. The 23 unique FAs are in bold.

S19. The list of significant function annotations for *g*C identified on the basis of *at least* two genes following IPA analysis on the gene list constructed using more stringent criteria. The table includes the category, general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the number of genes responsible for the enrichment signal and their gene names. The 8 unique FAs are in bold.

S20. The full IPA output when running analysis with all cell lines (excluding cancer cell lines) for *g*F**.** The table includes the category, the general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the number of genes responsible for the enrichment signal and their gene names.

S21. The full IPA output when running analysis with all cell lines (excluding cancer cell lines) for *g*C. The table includes the category, the general function, the specific function annotation (FA), the Fisher Exact Test (FET) P-value, the number of genes responsible for the enrichment signal and their gene names.

S22. All “IPA Ready” genes. Of the 34,109 Ensembl 54 genes tested, 19,275 were in the IPA database and thus available for pathway analysis. Also included are chromosome (CHR) and positional information (start and end, based on NCBI bld 36), gene symbol (if available), the gene description, the number of SNPs assigned to the gene after QC in the NCNG and in the CAGES.

**Supporting Information References**

Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., McGhee, K., Lopez, L., Gow, A.J., Corley, J., Redmond, P., Fox, H.C., Haggarty, P., Whalley, L.J., McNeill, G., Goddard, M.E., Espeseth, T., Lundervold, A.J., Reinvang, I., Pickles, A., Steen, V.M., Ollier, W., Porteous, D.J., Horan, M., Starr, J.M., Pendleton, N., Visscher, P.M. & Deary, I.J. (2011) Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol Psychiatry* **16,** 996-1005.

Delis, D.C., Karmer, J.H. & Kaplan, E. (2000) *California Verbal Learning Test*. 2nd edn. Psychological Corporation, San Antonio, TX.

Espeseth, T., Greenwood, P.M., Reinvang, I., Fjell, A.M., Walhovd, K.B., Westlye, L.T., Wehling, E., Lundervold, A., Rootwell, H. & Parasuraman, R. (2006) Interactive effects of APOE and CHRNA4 on attention and white matter volume in healthy middle-aged and older adults. *Cogn Affective Behav Neurosci* **6,** 31-43.

Trenerry, M.R. (1989) *Stroop Neuropsychological Screening Test Manual*. Psychological Assessment Resources, Odessa, FL.

Wechsler, D. (1999) *Wechsler Abbreviated Scale of Intelligence*. Psychological Corporation, San Antonio, TX.