bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: Gene Expression Correlates of the Cortical Network Underlying Sentence Processing

Abbreviated title: Genes & Sentence Processing Network

Authors: Xiang-Zhen Kong[1], Nathalie Tzourio-Mazoyer[2,3,4], Marc Joliot[2,3,4], Evelina Fedorenko[5,6,7], Jia Liu[8], Simon E. Fisher[1, 9], Clyde Francks[1, 9]

Affiliations:

1. Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands.

2. Univ. Bordeaux, GIN, IMN, UMR 5293, Bordeaux, France.

3. CNRS, GIN, IMN, UMR 5293, Bordeaux, France.

4. CEA, GIN, IMN, UMR 5293, Bordeaux, France.

5. McGovern Institute for Brain Research, MIT, Cambridge, MA, 02139, USA.

6. Department of Psychiatry, Harvard Medical School, Boston, MA, 02115, USA.

7. Department of Psychiatry, Massachusetts General Hospital, Boston, MA, 02114, USA.

8. Faculty of Psychology, Beijing Normal University, Beijing 100875, China.

9. Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands.

Correspondence Author:
Clyde Francks, Ph.D. (clyde.francks@mpi.nl)
Xiang-Zhen Kong, Ph.D. (xiangzhen.kong@outlook.com)

Number of Pages: 32; **Number of Figures/Tables:** 3/1.

Number of words for Abstract: 149**; for Introduction:** 1024**; for Discussion:** 2103**.**

Total number of words: 12863

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

A pivotal question in modern neuroscience is which genes regulate brain circuits that underlie cognitive functions. However, the field is still in its infancy. Here we report an integrated investigation of the high-level language network (i.e., sentence processing network) in the human cerebral cortex, combining regional gene expression profiles, task fMRI, and resting-state functional network approaches. We revealed reliable gene expression-functional network correlations using multiple datasets and network definition strategies, and identified a consensus set of genes related to connectivity within the sentence-processing network. The genes involved showed enrichment for neural development functions, as well as association signals with autism, which can involve disrupted language functioning. Our findings help elucidate the molecular basis of the brain's infrastructure for language. The integrative approach described here will be useful to study other complex cognitive traits.

Keywords: Sentence Processing Network, Language, Functional Network, Transcriptomic Network, Neural Development, Autism

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A pivotal question in modern neuroscience is which genes regulate brain circuits that underlie cognitive functions. In the past decade, imaging genetics has provided a powerful approach for exploring this question in humans, by combining neuroimaging data and genotype information from the same subjects and searching for associations between interindividual variability in neuroimaging phenotypes and genotypes within a sample. While numerous imaging genetics studies have now been published (e.g., [1, 2, 3]), there remain key issues which affect the field, including sample size limitations, the need to correct for multiple comparisons, and the small effect sizes that are typical of associations with common gene variants. Recently, researchers have begun to probe human gene-brain associations not only through genotypes, but also using gene expression profiles in brain tissues [4, 5, 6, 7, 8, 9]. This approach has brought important new advances: transcriptional profiles have been linked to neural architecture with respect to both functional connectivity measured during the resting state (also called intrinsic connectivity) [4, 5] and structural connectivity [4, 10], as well as to alterations of connectivity in brain disorders such as schizophrenia [6], autism spectrum disorder (ASD) [7], and Huntington's disease [8]. For example, in one study, Richiardi and colleagues (2015) used data from resting-state functional magnetic resonance imaging (rs-fMRI) to show that the network modularity structure derived from functional connectivity patterns across the cortex was correlated with inter-regional similarity of gene transcription profiles. Specifically, regions within a module (i.e., subnetwork) showed more similar gene expression profiles than across different modules. In another study, Romme et al. (2016) investigated the transcriptional profiles of a set of genes known to contain inherited variants associated with schizophrenia, and found that they were significantly correlated with regional reductions in the strength of white matter connections in patients with a schizophrenia diagnosis.

Past studies combining gene expression and brain imaging data have provided strong evidence that patterns of gene expression co-vary with anatomical and functional organization of the human brain. However, the ultimate goal is to gain a rich and detailed picture of the genetic and molecular mechanisms that support each core cognitive ability. Here we will focus on the quintessential and human-unique capacity of language. Previous studies have suggested that language-related cognitive performance is highly heritable (e.g., [11, 12, 13, 14]), and that brain activations associated with semantic comprehension tasks are also heritable [14]. Moreover, genetic factors also play a substantial role in susceptibility to language-related neurodevelopmental disorders such as childhood apraxia of speech [15], developmental language disorder (specific language impairment) and developmental dyslexia [16, 17]. Crucially, although a small number of genes – such as *FOXP2* (e.g., [18, 19, 20]) – have now been unambiguously linked to language-related disorders, these genes cannot by themselves explain the large majority of heritable variation, nor can they conceivably create or maintain the necessary brain circuits underlying language without interacting with a large number of other genes [21, 22].

In addition, linguistic deficits are often found with heritable, neurodevelopmental disorders for which impaired language function is not necessarily diagnostic, including intellectual disability,

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

autism spectrum disorder (ASD), and schizophrenia [23, 24, 25, 26, 27, 28, 29, 30, 31, 32]). Linguistic ability also correlates with intelligence in the general population [33]. Thus, identifying the molecular mechanisms and genes associated with language will not only i) yield a better understanding of the biological pathways that lead to the emergence of language phylo- and onto-genetically, but also ii) help identify susceptibility factors for language impairments in neuropsychiatric conditions, which could lead to improved diagnostic and treatment strategies.

To shed further light on the genetic and molecular architecture underpinning language circuits, here we synergistically combined task fMRI data, resting-state functional connectivity approaches, and gene transcription profiles in the human brain. Specifically, we targeted sentence-level processing as an essential, high-level linguistic function, which has been linked to a network of regions particularly in the left temporal and frontal cortices [34, 35, 36, 37, 38, 39, 40], as opposed to lower level language-related functions which can rely, for example, on primary auditory and motor areas. First, we defined the cortical regions for left hemispheric sentence processing based on task fMRI data, using three different sets of criteria and data to ensure robustness and generalizability across approaches. Then, we estimated the intrinsic functional connectivity networks among these regions, using rs-fMRI data from two independent datasets, and examined the correlations between these functional connectivity patterns and the corresponding inter-regional similarity patterns of gene expression. Next, we assessed the contributions of each individual gene to the observed correlations, and identified a consensus set of genes across all six analyses (i.e., three definition strategies for the sentence processing regions by two rs-fMRI datasets for estimating the functional connectivity). Finally, using several bioinformatics databases, we explored the biological roles, and expression specificity of these genes, and tested whether they showed an enrichment for association signals with ASD, schizophrenia or intelligence, using genome-wide association study (GWAS) data. We also analyzed three other functional networks by way of comparison to these language-related networks, which were the spatial navigation network, fronto-parietal multiple demand network, and default mode network.

Results

Fig. 1 shows a schematic of our approach for measuring the correlation between functional connectivity and gene expression profiles within a given network of brain regions. This analysis pipeline consisted of a) defining sets of cortical regions using task activation data (directly or via meta-analysis), b) estimating the resting-state functional connectivity and transcriptomic networks among these regions, and c) assessing the similarity between the functional connectivity and transcriptomic networks, along with subsequently d) estimating each gene's individual contribution (see STAR Methods). The details of each dataset and procedure are described in STAR Methods.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

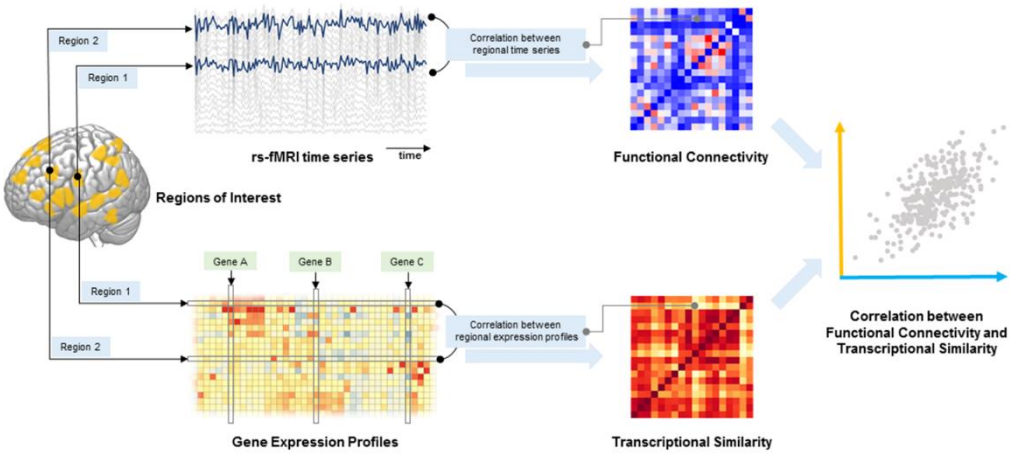


Fig. 1. Schematic of the pipeline for computing the correlation between resting state functional connectivity and transcriptomic similarity, within a network of regions first defined according to task fMRI data. Regional time series from rs-fMRI were used to estimate functional connectivity between each pair of regions (in the upper matrix, red indicates stronger functional connectivity; blue indicates weaker functional connectivity). Regional gene expression profiles were applied to estimate inter-regional transcriptional similarity (in the lower matrix, red indicates higher similarity; yellow indicates lower similarity; lower panel). Resting-state functional connectivity patterns (orange axis in the right-hand panel) are correlated with patterns of transcriptional similarity (blue axis in the right-hand panel).

Functional Networks

Given the absence of a universally agreed upon protocol for localizing brain regions that support high-level language processing [34], we used three different definition strategies: i) Supramodal Sentence Areas (SmSA) based on the concordance of activation across three language fMRI tasks and leftward lateralization completed by 144 healthy right-handers [35], ii) Synthesized Sentence Areas (SSA) based on large-scale neuroimaging meta-analysis of fMRI studies [41], and iii) One-contrast Sentence Areas (OcSA) based on the probabilistic activation map of a single language-task fMRI contrast [42] (see Methods). The three resulting maps of sentence processing network showed considerable overlap, and were consistent with previous studies, especially as regards core language regions such as the temporal and frontal regions (Fig. 2; Supplemental Table S1) [34]. Regions were defined according to the AICHA brain atlas, which is derived from rs-fMRI connectivity data, with each region showing homogeneity of functional temporal activity within itself [43].

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

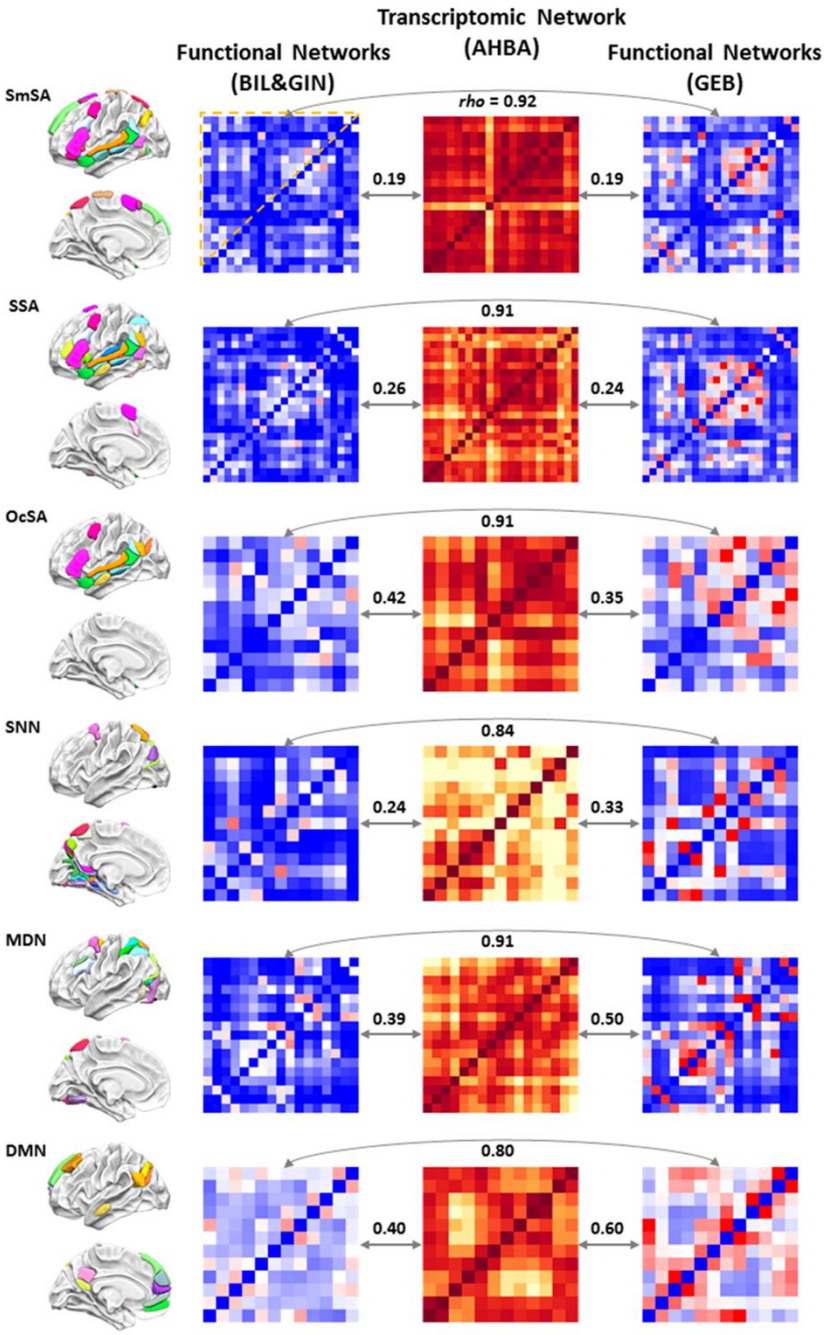


Fig. 2. Functional and transcriptomic networks of the present study. First (left-side) column: the sets of cortical regions defined based on task fMRI data: top three (SmSA, SSA, and OcSA) correspond to the three different ways of defining the network for sentence processing, and bottom three correspond to the three other comparison networks. Second and fourth columns: the functional connectivity networks based on rs-fMRI from the BIL&GIN dataset and GEB dataset respectively. Third column: the transcriptomic networks based on the AHBA database. Within each matrix, the colors indicate the strength of pairwise inter-regional correlations (from 0 in blue, to 1 in red) or transcriptomic inter-regional similarity (from 0.9 in yellow to 1.0 in red). Correlation values among the right three columns are the Spearman correlations (*rho*) of each pair of matrices for a given network, based on the upper triangle part of each matrix. All correlations were significant (p <0.01). SmSA, Supramodal Sentence Areas; SSA, Synthesized Sentence Areas; OcSA, One-contrast Sentence Areas; SNN, Spatial Navigation Network; MDN, Multiple Demands Network; DMN, Default Mode Network.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To compare to language-related networks, we also analyzed three other functional cortical networks: the spatial navigation network (SNN), fronto-parietal multiple demand network (MDN), and default mode network (DMN). The sets of cortical regions defined for these networks (see Methods) appeared consistent with previous literature [44, 45, 46] and showed little overlap with the sentence processing networks defined above (Fig. 2). More information about the spatial distribution of each definition and their overlaps can be seen in Supplemental Table S1.

Connectivity within a given functionally-defined set of regions was estimated based on inter-regional synchronization of rs-fMRI time courses. The connectivity patterns based on two independent rs-fMRI datasets (BIL&GIN and GEB; see Methods) were highly reproducible for all functional networks (*rho* > 0.80; Fig. 2). Corresponding transcriptomic networks for each functionally defined set of regions were calculated based on post-mortem cortical gene expression data, and using pairwise similarities of regional gene expression profiles (Methods). These analyses were restricted to the top 5% of all genes (i.e., 867 genes) which showed the highest differential stability of their expression levels across donors, in cerebral cortical data from the Allen Brain Atlas [47]. Within a given network, different pairs of regions varied in how similar they were in overall gene expression, although all regional pairwise correlations were high (greater than 0.9) (Fig. 2).

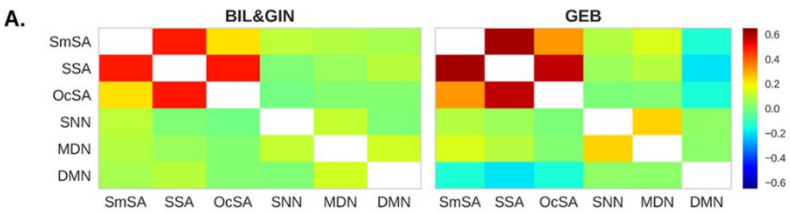
Similarity between Functional and Transcriptomic Networks

We used correlation analysis to test whether regions with more similar gene transcription profiles show stronger resting-state functional connectivity, within each specific functional network. As expected, based on data from the BIL&GIN dataset, we found significant correlations between functional connectivity and the corresponding transcriptomic networks for each sentence processing network definition (Fig. 2; SmSA: *rho* = 0.19, *p* = 0.0048; SSA: *rho* = 0.26, *p* < 0.0001; OcSA: *rho* = 0.42, *p* = 0.00045). In addition, we obtained similar results for each of the comparison networks (Fig. 2; SNN: *rho* = 0.24, *p* = 0.0020; MDN: *rho* = 0.39, *p* < 0.0001; and DMN: *rho* = 0.40, *p* = 0.00081). Using connectivity data from the independent rs-fMRI dataset GEB, highly similar results were found (Fig. 2; SmSA: *rho* = 0.19, *p* = 0.0054; SSA: *rho* = 0.24, *p* = 0.00018; OcSA: *rho* = 0.35, *p* = 0.0042; SNN: *rho* = 0.33, *p* < 0.0001; MDN: *rho* = 0.50, *p* < 0.0001; and DMN: *rho* = 0.60, *p* < 0.0001). We also found similar correlations when applying a more inclusive threshold for gene expression differential stability across donors (stability > 0.25, i.e. the top 10% genes in stability) (Supplemental Table S2). As a negative control, we also tested the bottom genes in differential stability across donors (i.e., bottom 5% or 10% of genes), and saw little evidence of significant gene-brain correlation (Supplemental Table S2), as expected. In addition, as spatial proximity may influence the estimation of both functional connectivity and transcriptomic networks ([48], see also [49]), we confirmed all correlations after controlling for the spatial distance (i.e., Euclidean distance) between centers of regions (all *p*s < 0.01) (Supplemental Table S3).

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene Contribution Index (GCI)

With a 'leave-one-out' procedure, we obtained 'gene contribution index' (GCI) scores for all individual genes, and for each functional network definition, which indicated the extent to which each gene affected the overall connectivity-transcriptome correlation for a given network (see Methods). To investigate the similarity of gene contribution patterns across different networks, we calculated the correlation between the GCI scores of each pair of networks. As expected, for the three different definitions of the sentence processing network, which involved largely overlapping sets of cortical regions, the GCI scores showed substantial correlations, which were also reproducible across the two independent rs-fMRI datasets (Fig. 3A; BIL&GIN: Mean r = 0.40, from 0.22 to 0.50, N = 3; GEB: Mean CGI r = 0.50, from 0.32 to 0.61, N = 3). However, the GCI scores showed low correlations across different functional networks (i.e., between each of the comparison networks with one another, and with any definition of the sentence processing network), which was again consistent across the two independent rs-fMRI datasets for measuring functional connectivity (Fig. 3A; BIL&GIN: Mean r = 0.058±0.047, N = 12; GEB: Mean r = 0.017±0.12, N = 12).



B.

ANK1	*DIRAS3*	*IGSF22*	*MET*	***SPON2***
ARHGAP25	*ELMO3*	*IQCJ*	***MGP***	*TGFBI*
ASGR2	*FAM162B*	*KCNS2*	*NT5M*	*TRPC6*
C12orf23	***FAM65B***	*KIAA1875*	*PGM2L1*	*VILL*
CBLN2	*FRAT1*	*LAIR2*	*PRRX1*	*ZYX*
CDR2L	*GLCCI1*	*LCP2*	*ROBO1*	
CNTN6	*GLDC*	*LOC145837*	***SERPINF1***	
COX7A1	*GPRIN1*	***LY6H***	*SHD*	
CTXN3	*HR*	*MAGED4B*	*SLC22A9*	

Fig. 3. Contributions of individual genes to overall gene-brain associations. (A) Correlations between GCI scores derived for each functional network, based on the two independent datasets BIL&GIN (left) and GEB (right). The three definitions of the sentence processing network showed relatively high correlations of their GCI scores, while different functional networks showed low correlations of their GCI scores. (B) The consensus set of genes (N = 41) for the six analyses of the sentence processing network (i.e., three functional definition strategies by two rs-fMRI datasets). Red color indicates genes that have been linked to language-related phenotypes in previous studies (e.g., reading deficits, intellectual disability, and autism), and genes **in bold** indicate significant differential expression within the sentence processing network regions compared to other cortical regions (FDR corrected p < 0.05). SmSA, Multimodal Sentence Areas; SSA, Synthesized Sentence Areas; OcSA, One-contrast Sentence Areas; SNN, Spatial Navigation Network; MDN, Multiple Demand Network; DMN, Default Mode network.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In order to derive a set of genes related to the sentence processing network with highest consistency, we identified 41 "consensus genes" which had positive CGI scores in all six analyses of this network, i.e. the three definition strategies (SmSA, SSA, and OcSA), by the two independent rs-fMRI datasets (BIL&GIN and GEB) (Fig. 3B). Several of these genes, including *ROBO1*, *MET*, *PRRX1*, *CNTN6*, and *CTXN3*, have been reported to affect language- or reading-related phenotypes such as dyslexia, and/or disorders that are often accompanied by linguistic impairments, i.e. intellectual disability, ASD, and schizophrenia (see Discussion). Further information on the consensus gene set is provided in Supplemental Table S4.

Biological Roles of Consensus Genes Associated with the Sentence Processing Network

Gene ontology analysis of the consensus set of 41 genes associated with the sentence processing network identified significant enrichment for terms mostly related to neural development, or likely to influence axon growth (Table 1). For example, five of the consensus gene set are members of the set 'axon guidance' ($p = 0.0029$), i.e. *ANK1*, *MET*, *ROBO1*, *TRPC6*, and *CNTN6*. A set of 5 genes contributed to terms related to actin cytoskeleton organization (ps <0.05), which also plays a role in regulating the extension and direction of axon growth (Coles and Bradke, 2015). Another significant enrichment was for the set 'neuron projection development' ($p = 0.012$), and there were 6 genes that drove the enrichment results, which were *SERPINF1*, *PRRX1*, *ROBO1*, *TRPC6*, *SPON2*, and *GPRIN1* (Table 1). To examine expression of these genes across human brain development, we used data from the Allen Institute's BrainSpan project, which includes human brain tissues from embryonic stages to adulthood, measured using RNA-sequencing (Miller et al., 2014). Each of these genes has detectable expression in the frontal and temporal regions during fetal development and in early childhood (Fig. 4). Some of the genes increase in expression through development all the way to adulthood, such as *MET*, *SERPINF1* and *PRRX1*. Others decrease such as *ROBO1*, *GPRIN1* and *TRPC6*, while still remaining expressed through to adulthood.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. Gene ontology terms for which the consensus set associated with the sentence processing network shows enrichment. BP, biological process; REAC, biological pathways based on Reactome.

Term names	Corrected p	# of term genes	Overlap	Intersecting genes
BP: neuron projection development	0.0118	902	6	*SERPINF1, PRRX1, ROBO1, TRPC6, SPON2, GPRIN1*
BP: actin filament-based process	0.0455	705	5	*MET, ZYX, ARHGAP25, VILL, IGSF22*
BP: actin cytoskeleton organization	0.0228	609	5	*MET, ZYX, ARHGAP25, VILL, IGSF22*
BP: supramolecular fiber organization	0.0196	590	5	*MET, ZYX, ARHGAP25, VILL, IGSF22*
BP: actin filament organization	0.00183	360	5	*MET, ZYX, ARHGAP25, VILL, IGSF22*
REAC: Axon guidance	0.00293	527	5	*ANK1, MET, ROBO1, TRPC6, CNTN6*

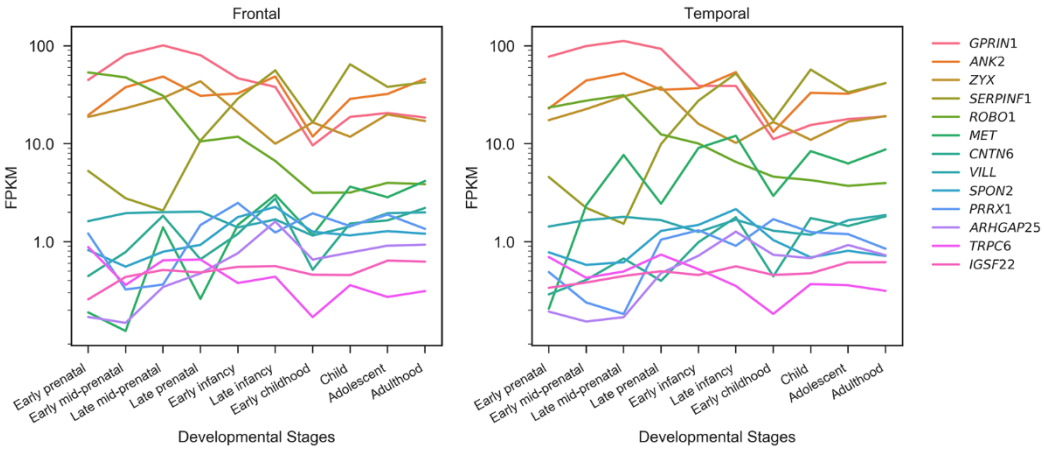


Fig. 4. Gene expression in human frontal and temporal cortices, from embryo (8 week post conception) to adulthood (40 years), from the BrainSpan atlas. Genes are listed in the legend in order of their averaged expression across all samples. The y axis uses a $\log_{10}$ scale to visualize change over time for each gene. The samples were dissected according to the BrainSpan Technical White Paper (http://www.brainspan.org/).

We found that 14 genes of the 41 consensus gene set showed differential expression when contrasting regions defined as belonging to the sentence processing network against those in the comparison networks (see Methods; uncorrected $p < 0.05$), among which 6 genes survived correction for multiple comparisons (FDR corrected $p < 0.05$): *C12orf23*, *FAM65B*, *LY6H*, *MGP*, *SERPINF1*, and *SPON2*, all with higher expression in the regions assigned to the sentence processing network. Notably, three of the six genes that were related to "neural projection development" in gene ontology analysis showed significantly higher expression within the sentence-processing regions, which were *SERPINF1*, *SPON2*, and *GPRIN1* (*p*s < 0.05; e.g. *SERPINF1*$t(46) = 3.85$, $p = 0.00036$; 95%

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

confidence interval of the difference: 0.24 to 0.76), while none showed lower expression in the sentence processing regions (*p*s >0.05). More information can be found in Supplemental Table S5.

Of the 41 consensus genes associated with the sentence processing network, data on 33 were available from a single-cell RNA-sequencing study of adult mouse cerebral cortex (Zhang et al. 2014). Among these 33 genes, a majority (i.e., 25) showed expression in one cell type that was at least 1.5 times higher than all other cell types, although not predominantly in neurons versus other cell types (Supplemental Table S6). For example, *MET* and *CTXN3* showed 3.57- and 1.53-fold expression in neurons compared to the maximum expression in other cell types, respectively. Data on seven of the 41 genes were available in another single-cell gene expression database [50], and two of these showed enrichment in interneurons: *ANK1* and *SHD*. Thus, among the five genes driving the significant enrichment result for the term 'axon guidance', *ANK1*, *MET*, *ROBO1*, and *TRPC6* had their highest levels in neurons. In addition, among the six genes driving the gene ontology enrichment results related to neural projection development, four showed their highest expressions in neurons, including *ROBO1*, *TRPC6*, *SPON2*, and *GPRIN1*.

Association with ASD, schizophrenia and intelligence

We were interested to test whether genes involved in the cortical language network also contain polymorphisms in the population which affect human cognitive or behavioural variation, or susceptibility to neuropsychiatric disorders. This analysis requires genome-wide association scan (GWAS) results from large-scale studies. No studies comprising more than 10,000 subjects have yet been published for reading/language measures in the general population, nor for disorders such as dyslexia or language impairment which involve language-related deficits. We analyzed ASD and schizophrenia, as these disorders can involve linguistic deficits, as well as intelligence in the general population, which correlates with linguistic abilities (see Introduction). Using GWAS summary statistics for ASD based on up to 7387 cases and 8567 controls [51], we found that the 41 consensus genes associated with the sentence processing network were significantly enriched for single nucleotide polymorphisms (SNPs) showing association with ASD (*beta* = 0.32, $p = 0.0080$). No such signal in relation to ASD was seen for the top genes ($N = 41$) with highest CGI scores for each of the comparison networks (SNN: $p = 0.45$; MDN: $p = 0.65$; DMN: $p = 0.074$). No significant enrichment was found for any functional network in relation to schizophrenia (GWAS based on up to 36,989 cases and 113,075 controls; [32]) (*p*s >0.50). For intelligence (GWAS based on 78,308 individuals; [52]), there was a significant enrichment for the top genes associated with the MDN (*beta* = 0.30, $p = 0.035$), with no other enrichment found (*p*s >0.30).

Discussion

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In this study, we combined gene transcription profiles in the human brain with task and resting-state fMRI data, and investigated the gene expression correlates of high-level linguistic network. Specifically, with six analyses based on complementary strategies and independent datasets, we revealed a significant correlation between the pattern of functional connectivity within the sentence processing network and the corresponding pattern of inter-regional gene expression similarity. To our knowledge, this is the first evidence for a link between gene transcription profiles and language networks. While some previous studies have suggested that transcription profiles are linked to patterns of structural and functional connectivity across the brain [4, 5, 10], this relationship could be driven by broad differences in gene expression between sensory and higher-order association cortices. Here, we focus on one core human cognitive ability – language – and examine a fine-grained pattern of brain-gene relationships within specific networks that support sentence processing. Across three definitions of sentence processing regions and using two independent resting-state fMRI datasets, we established and characterized a relationship with gene expression: pairs of language-sensitive brain regions that show synchronization during rest also show more similar profiles of gene expression. The underlying basis of this association is unknown. One possibility is that functionally linked cortical regions are more likely to share specific aspects of neuronal physiology and developmental trajectory, which support their temporally synchronized activity.

We identified a consensus set of genes most consistently linked to the transcriptome-connectivity correlation within the sentence processing network, and thereby gained new insights into the molecular bases of our language-ready brain. This gene set was enriched for functions including axon guidance and actin cytoskeleton organization, and the genes driving this enrichment included *MET*, *ROBO1*, *ANK1*, and *TRPC6*, which showed their highest expression levels in neurons in single-cell gene expression data from mouse cerebral cortex. The consensus genes associated with language-related networks were also enriched for neuron projection development. Each of the six genes driving this enrichment, i.e. *SERPINF1*, *PRRX1*, *ROBO1*, *TRPC6*, *SPON2*, and *GPRIN1*, has detectable expression in frontal and temporal regions during fetal development and early childhood, and might play important roles during language network development. The fact that these genes, known for their neurodevelopmental roles, are expressed in adult cerebral cortex in a manner linked to functional connectivity within language networks, also suggests that they have continuing roles in maintaining adult cortical circuitry for its regionally-specialized roles. Further support for the particular importance of these genes for the sentence processing network came from the fact that most of them showed higher gene expression in language-related regions than elsewhere in the cortex, and none showed significantly lower expression in the language-related regions.

Several of the consensus genes associated with the sentence processing network have previously been reported to contain common polymorphisms or rare mutations which impact on language- and reading-related phenotypes, or else disorders which can be accompanied by reduced linguistic abilities

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such as intellectual disability, ASD or schizophrenia. *ROBO1* has well established roles in brain development [53, 54], and has been implicated in dyslexia as well as phonological short term memory [54, 55], while its homologue *ROBO2* has been linked with expressive vocabulary during early language acquisition [53]. Mutations in *MET* are a risk factor for ASD [56, 57, 58, 59], and *MET* is also regulated by the transcription factor *FOXP2* [56], which causes developmental verbal dyspraxia when mutated [18]. *PRRX1* is associated with intellectual disability and delayed language acquisition [60]. Mutations of *CNTN6* have been reported in patients with speech and language delays, intellectual disability, and atypical ASD [61]. *CTXN3* has been linked to schizophrenia [62, 63]. These observations suggest that our consensus gene set linked to the sentence processing network might provide additional candidates for future studies of language-related individual differences, and neurodevelopmental disorders.

No GWAS studies comprising more than 10,000 subjects have yet been published for reading/language measures in the general population, nor for disorders such as dyslexia or language impairment which involve language-related deficits. However, through analysis of large-scale GWAS summary statistics for ASD, a disorder that can also involve linguistic deficits [23, 24], we found that the consensus gene set associated with connectivity in the sentence processing network is enriched for common SNPs that contribute to the polygenic liability to ASD in the population. It is therefore possible that variants in these genes associate with ASD due, at least in part, to dysfunction of high-level language networks. Note that, while language-related deficits may no longer be considered core to ASD in the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM 5) [64], the earlier DSM-IV criteria were used for the large-scale GWAS whose results we used in the present study [51], in which language delays were considered an important aspect of the disorder [65]. No such enrichment signal for ASD was observed for gene sets associated with the other functional networks that we analyzed by way of comparison, i.e., spatial navigation, multiple demand, and default mode networks. Interestingly, a recent large-scale brain imaging analysis of ASD found that cortical thinning was present in many of the regions included in our definitions of the sentence-processing network [66]. For schizophrenia and intelligence we found no enrichment of association signals within the consensus gene set associated with the sentence processing network. This pattern may relate to severity, as language deficits in schizophrenia patients have been reported to be less severe than in ASD patients (again based on earlier diagnostic criteria) [67, 68]. Alternatively the genetic underpinnings of high-level language processing may be of little relevance to schizophrenia, or intelligence in the general population. However, statistical power to detect these relations may also be an issue.

There is at present no consensus in the field on what should constitute the precise high-level language network and how best to define it [34, 35]. To circumvent this problem, we used three complementary approaches for defining brain regions important for sentence-level processing: one based on a conjunction of three task contrasts and functional laterality (SmSA; [35]), one based on a large-scale meta-analysis of prior neuroimaging studies (SSA; [41]), and one based on a single task

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

contrast (OcSA; [42]). Reassuringly, the three approaches yielded similar definitions of high-level language networks, especially with respect to areas in the inferior frontal and middle temporal gyri. Likely because of the overlap between the three network definitions, their transcriptomic correlates were also similar, as well as the contributions of individual genes. This overall concordance supports the validity of the different definitions, allowing us to define a consensus set of genes that emerged consistently across all definitions and both of the rs-fMRI datasets, and is thus not dependent solely on any individual approach for defining the functional network. The set of regions for sentence processing defined here could be used in future studies of vast datasets, such as the UK Biobank [69] or ENIGMA Consortium [70], which include neuroanatomical and/or intrinsic connectivity data but limited task fMRI data for language functions. For instance, the functional regions defined here can be used to investigate structural/functional variability during maturation, aging, and/or pathological processes in the general population, including those associated with developmental language disorders.

For the comparison networks included in this study, i.e. the spatial navigation, multiple demand, and default mode networks, we also found similar overall correlations between the functional connectivity networks and the corresponding transcriptomic networks, although with largely different individual genes contributing. For example, *MET* showed a relatively high contribution to the connectivity-transcriptomic association for the sentence processing network, but its contribution to other functional networks was negligible. Therefore, although the existence of connectivity-transcriptome correlations appears to be ubiquitous across functional networks, our study yields the important basic insight that different functional networks can involve differently weighted genetic contributions at the level of cortical gene expression. This is broadly consistent with genetic correlations between different cognitive abilities such as verbal and non-verbal performance, as assessed in population genetic analysis, which often indicate shared but also independent genetic effects on such pairs of traits [71]. Our study therefore suggests a novel approach to complement existing genetic epidemiological approaches, for understanding the general versus specific influences on diverse cognitive abilities. Note that, in the present study, in order to achieve comparable statistical power in the network similarity analysis across different networks, we purposely defined similar numbers of top regions for each comparison network, even though this meant using different thresholds for including regions in each comparison network. Further studies may investigate how using more or less inclusive definitions of these functional networks affect relationships with gene expression.

The present study focused on only a subset of genes with relatively high stability across a small number of donors in the Allen Brain database. Post-mortem brain tissues suitable for transcriptomic analysis are difficult to collect from individuals who were healthy immediately prior to death, which is necessary since RNA degrades within hours after death. Thus the availability of high-quality gene expression data from the human brain is necessarily limited. Most of the Allen brain data, and all the

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data used for the present study, are based on the older and relatively noisy transcriptomic technology of microarrays, rather than the more accurate, latest method of RNA sequencing. Some genes known to be involved in language, such as *FOXP2* [18], had to be excluded from our analyses because of the inclusion criterion of stability across donors. *FOXP2* showed low inter-donor differential stability of 0.22 across cerebral cortex samples, especially low across frontal samples (-0.07) and across temporal samples (0.16) in the Allen brain data [47]. Thus, it is likely that data of sufficient quality were not available for other genes too, that might have been of relevance to language networks, such that future studies using RNA sequencing in larger numbers of individuals, and with more sampling per cortical region, would be well motivated. In addition, inter-individual variabilities have been observed in the precise locations of language-sensitive regions [42, 72] as well as the cytoarchitectonic features of higher level cognitive areas (e.g., BA 44 and 45) [73]. Gene expression data from a larger number of individuals, ideally when both anatomical and functional data from the same individuals are available, would help to improve upon this aspect.

A further limitation of our study is the use of transcriptomic data from blocks of cerebral cortical tissues, which comprised many cell types. Future databases based on single-cell transcriptomics would likely provide further insights. Although efforts to produce such data are underway for a limited number of human cerebral cortical regions, no database currently exists which has broad mapping over the cerebral cortex. It will be a major undertaking in the future for the human brain science community to achieve a widespread cerebral cortical gene expression map at single cell resolution. For the time being, we queried our consensus genes associated with the language-related network using single cell transcriptomic data from the adult mouse cortex, which gave information about whether specific genes are relatively more highly expressed in neurons versus some major classes of glial cells in the mouse. However, even if a gene of interest were expressed at comparable or higher levels in glia than neurons, it might still influence neuronal physiology and circuit properties, either directly through its expression in neurons, or indirectly via the interactions of surrounding glial cells with neurons.

In sum, we provide a first description of the overall transcriptomic correlates of brain networks underlying high-level linguistic processing, as well as identifying a set of individual genes likely to be most important. These findings help elucidate the molecular basis of language networks, as distinct from functional networks important for other aspects of cognition. A link of this genetic infrastructure to ASD is also suggested by our data. Finally, we propose functional connectivity and gene expression analysis as a complementary approach to existing genetic epidemiological and genetic association approaches, for understanding complex cognitive traits.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Hariri AR, Drabant EM, Weinberger DR. Imaging genetics: Perspectives from studies of genetically driven variation in serotonin function and corticolimbic affective processing. *Biological Psychiatry* 2006, **59**(10)**:** 888-897.

2. Bigos KL, Weinberger DR. Imaging genetics-days of future past. *Neuroimage* 2010, **53**(3)**:** 804-809.

3. Hibar DP, Adams HHH, Jahanshad N, Chauhan G, Stein JL, Hofer E*, et al.* Novel genetic loci associated with hippocampal volume. *Nat Commun* 2017, **8:** 13624.

4. Richiardi J, Altmann A, Milazzo AC, Chang C, Chakravarty MM, Banaschewski T*, et al.* BRAIN NETWORKS. Correlated gene expression supports synchronous activity in brain networks. *Science* 2015, **348**(6240)**:** 1241-1244.

5. Vertes PE, Rittman T, Whitaker KJ, Romero-Garcia R, Vasa F, Kitzbichler MG*, et al.* Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos Trans R Soc Lond B Biol Sci* 2016, **371**(1705).

6. Romme IA, de Reus MA, Ophoff RA, Kahn RS, van den Heuvel MP. Connectome Disconnectivity and Cortical Gene Expression in Patients With Schizophrenia. *Biol Psychiatry* 2017, **81**(6)**:** 495-502.

7. Romero-Garcia R, Warrier V, Bullmore ET, Baron-Cohen S, Bethlehem RAI. Synaptic and transcriptionally downregulated genes are associated with cortical thickness differences in autism. *Mol Psychiatry* 2018.

8. McColgan P, Gregory S, Seunarine KK, Razi A, Papoutsi M, Johnson E*, et al.* Brain Regions Showing White Matter Loss in Huntington's Disease Are Enriched for Synaptic and Metabolic Genes. *Biol Psychiatry* 2018, **83**(5)**:** 456-465.

9. Kong XZ, Song Y, Zhen Z, Liu J. Genetic Variation in S100B Modulates Neural Processing of Visual Scenes in Han Chinese. *Cereb Cortex* 2017, **27**(2)**:** 1326-1336.

10. Fulcher BD, Fornito A. A transcriptional signature of hub connectivity in the mouse connectome. *Proc Natl Acad Sci U S A* 2016, **113**(5)**:** 1435-1440.

11. Dale PS, Simonoff E, Bishop DVM, Eley TC, Oliver B, Price TS*, et al.* Genetic influence on language delay in two-year-old children. *Nature Neuroscience* 1998, **1**(4)**:** 324-328.

12. Newbury DF, Bishop DVM, Monaco AP. Genetic influences on language impairment and phonological short-term memory. *Trends in Cognitive Sciences* 2005, **9**(11)**:** 528-534.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

13. Thompson PM, Cannon TD, Narr KL, van Erp T, Poutanen VP, Huttunen M*, et al.* Genetic influences on brain structure. *Nature Neuroscience* 2001, **4**(12)**:** 1253-1258.

14. Guen YL, Amalric M, Pinel P, Pallier C, Frouin V. Shared genetic aetiology between cognitive performance and brain activations in language and math tasks. *bioRxiv* 2018.

15. Eising E, Carrion-Castillo A, Vino A, Strand EA, Jakielski KJ, Scerri TS*, et al.* A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry* 2018.

16. Ramus F, Marshall CR, Rosen S, van der Lely HKJ. Phonological deficits in specific language impairment and developmental dyslexia: towards a multidimensional model. *Brain* 2013, **136:** 630-645.

17. Deriziotis P, Fisher SE. Speech and Language: Translating the Genome. *Trends Genet* 2017, **33**(9)**:** 642-656.

18. Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* 2001, **413**(6855)**:** 519-523.

19. Morgan A, Fisher SE, Scheffer I, Hildebrand M. FOXP2-Related Speech and Language Disorders. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Stephens K*, et al.* (eds). *GeneReviews((R))*: Seattle (WA), 1993.

20. Fisher SE, Scharff C. FOXP2 as a molecular window into speech and language. *Trends Genet* 2009, **25**(4)**:** 166-177.

21. Graham SA, Fisher SE. Understanding Language from a Genomic Perspective. *Annual review of genetics* 2015, **49:** 131-160.

22. Konopka G, Roberts TF. Insights into the Neural and Genetic Basis of Vocal Communication. *Cell* 2016, **164**(6)**:** 1269-1276.

23. Kleinhans NM, Muller RA, Cohen DN, Courchesne E. Atypical functional lateralization of language in autism spectrum disorders. *Brain Research* 2008, **1221:** 115-125.

24. Lombardo MV, Pierce K, Eyler LT, Barnes CC, Ahrens-Barbeau C, Solso S*, et al.* Different Functional Neural Substrates for Good and Poor Language Outcome in Autism. *Neuron* 2015, **86**(2)**:** 567-577.

25. Tomblin B. Co-morbidity of autism and SLI: kinds, kin and complexity. *Int J Lang Commun Disord* 2011, **46**(2)**:** 127-137.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. Kaufman L, Ayub M, Vincent JB. The genetic basis of non-syndromic intellectual disability: a review. *J Neurodev Disord* 2010, **2**(4)**:** 182-209.

27. Bearden CE, Rosso IM, Hollister JM, Sanchez LE, Hadley T, Cannon TD. A prospective cohort study of childhood behavioral deviance and language abnormalities as predictors of adult schizophrenia. *Schizophr Bull* 2000, **26**(2)**:** 395-410.

28. Tager‐Flusberg H, Paul R, Lord C. Language and Communication in Autism. In: Volkmar FR, Paul R, Klin A, Cohen D (eds). *Handbook of Autism and Pervasive Developmental Disorders*, vol. 1. Wiley: New York, 2005.

29. Geschwind DH, Flint J. Genetics and genomics of psychiatric disease. *Science* 2015, **349**(6255)**:** 1489-1494.

30. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D*, et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 2014, **515**(7526)**:** 216-221.

31. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S*, et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* 2008, **455**(7210)**:** 232-236.

32. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014, **511**(7510)**:** 421-427.

33. Silva PA, Williams S, Mcgee R. A Longitudinal-Study of Children with Developmental Language Delay at Age 3 - Later Intelligence, Reading and Behavior Problems. *Developmental Medicine and Child Neurology* 1987, **29**(5)**:** 630-640.

34. Fedorenko E, Thompson-Schill SL. Reworking the language network. *Trends in Cognitive Sciences* 2014, **18**(3)**:** 120-126.

35. Labache L, Joliot M, Saracco J, Jobard G, Hesling I, Zago L*, et al.* FALCON: a Functional Atlas of Language COmprehension Networks based on multiple task-induced activation mapping and graph analysis of intrinsic connectivity in 137 healthy right-handers. *BioRxiv* 2018.

36. Vigneau M, Beaucousin V, Herve PY, Duffau H, Crivello F, Houde O*, et al.* Meta-analyzing left hemisphere language areas: Phonology, semantics, and sentence processing. *Neuroimage* 2006, **30**(4)**:** 1414-1432.

37. Price CJ. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 2012, **62**(2)**:** 816-847.

38. Price CJ. The anatomy of language: a review of 100 fMRI studies published in 2009. *Ann N Y Acad Sci* 2010, **1191:** 62-88.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

39. Dronkers NF, Wilkins DP, Van Valin RD, Jr., Redfern BB, Jaeger JJ. Lesion analysis of the brain areas involved in language comprehension. *Cognition* 2004, **92**(1-2)**:** 145-177.

40. Fedorenko E, Behr MK, Kanwisher N. Functional specificity for high-level linguistic processing in the human brain. *Proc Natl Acad Sci U S A* 2011, **108**(39)**:** 16428-16433.

41. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods* 2011, **8**(8)**:** 665-670.

42. Fedorenko E, Hsieh PJ, Nieto-Castanon A, Whitfield-Gabrieli S, Kanwisher N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J Neurophysiol* 2010, **104**(2)**:** 1177-1194.

43. Joliot M, Jobard G, Naveau M, Delcroix N, Petit L, Zago L*, et al.* AICHA: An atlas of intrinsic connectivity of homotopic areas. *J Neurosci Methods* 2015, **254:** 46-59.

44. Kong XZ, Wang X, Pu Y, Huang L, Hao X, Zhen Z*, et al.* Human navigation network: the intrinsic functional organization and behavioral relevance. *Brain Struct Funct* 2017, **222**(2)**:** 749-764.

45. Buckner RL, Andrews-Hanna JR, Schacter DL. The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci* 2008, **1124:** 1-38.

46. Crittenden BM, Mitchell DJ, Duncan J. Task Encoding across the Multiple Demand Cortex Is Consistent with a Frontoparietal and Cingulo-Opercular Dual Networks Distinction. *J Neurosci* 2016, **36**(23)**:** 6147-6155.

47. Hawrylycz M, Miller JA, Menon V, Feng D, Dolbeare T, Guillozet-Bongaarts AL*, et al.* Canonical genetic signatures of the adult human brain. *Nat Neurosci* 2015, **18**(12)**:** 1832-1844.

48. Pantazatos SP, Li XY. Commentary: BRAIN NETWORKS. Correlated Gene Expression Supports Synchronous Activity in Brain Networks. Science 348, 1241-4. *Front Neurosci-Switz* 2017, **11**.

49. Richiardi J, Altmann A, Greicius M. Distance Is Not Everything In Imaging Genomics Of Functional Networks: Reply To A Commentary On Correlated Gene Expression Supports Synchronous Activity In Brain Networks. *bioRxiv* 2017.

50. Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A*, et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015, **347**(6226)**:** 1138-1142.

51. Autism Spectrum Disorders Working Group of The Psychiatric Genomics C. 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* 2017, **8:** 21.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E*, et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat Genet* 2017, **49**(7)**:** 1107-1112.

53. St Pourcain B, Cents RA, Whitehouse AJ, Haworth CM, Davis OS, O'Reilly PF*, et al.* Common variation near ROBO2 is associated with expressive vocabulary in infancy. *Nat Commun* 2014, **5:** 4831.

54. Hannula-Jouppi K, Kaminen-Ahola N, Taipale M, Eklund R, Nopola-Hemmi J, Kaariainen H*, et al.* The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *Plos Genetics* 2005, **1**(4)**:** 467-474.

55. Bates TC, Luciano M, Medland SE, Montgomery GW, Wright MJ, Martin NG. Genetic Variance in a Component of the Language Acquisition Device: ROBO1 Polymorphisms Associated with Phonological Buffer Deficits. *Behavior Genetics* 2011, **41**(1)**:** 50-57.

56. Mukamel Z, Konopka G, Wexler E, Osborn GE, Dong H, Bergman MY*, et al.* Regulation of MET by FOXP2, genes implicated in higher cognitive dysfunction and autism risk. *J Neurosci* 2011, **31**(32)**:** 11437-11442.

57. Sousa I, Clark TG, Toma C, Kobayashi K, Choma M, Holt R*, et al.* MET and autism susceptibility: family and case-control studies. *Eur J Hum Genet* 2009, **17**(6)**:** 749-758.

58. Jackson PB, Boccuto L, Skinner C, Collins JS, Neri G, Gurrieri F*, et al.* Further evidence that the rs1858830 C variant in the promoter region of the MET gene is associated with autistic disorder. *Autism Res* 2009, **2**(4)**:** 232-236.

59. Campbell DB, Sutcliffe JS, Ebert PJ, Militerni R, Bravaccio C, Trillo S*, et al.* A genetic variant that disrupts MET transcription is associated with autism. *Proc Natl Acad Sci U S A* 2006, **103**(45)**:** 16834-16839.

60. Lam F, Morris C. Nine Year Old Boy with Chromosome 1q23.3-q25.1 Deletion. *American Journal of Medical Genetics Part A* 2016, **170**(11)**:** 3013-3017.

61. Kashevarova AA, Nazarenko LP, Schultz-Pedersen S, Skryabin NA, Salyukova OA, Chechetkina NN*, et al.* Single gene microdeletions and microduplication of 3p26.3 in three unrelated families: CNTN6 as a new candidate gene for intellectual disability. *Mol Cytogenet* 2014, **7**.

62. Potkin SG, Macciardi F, Guffanti G, Fallon JH, Wang Q, Turner JA*, et al.* Identifying gene regulatory networks in schizophrenia. *Neuroimage* 2010, **53**(3)**:** 839-847.

63. Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I*, et al.* Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am J Hum Genet* 2003, **73**(1)**:** 34-48.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

64. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

65. Grzadzinski R, Huerta M, Lord C. DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. *Mol Autism* 2013, **4**(1)**:** 12.

66. van Rooij D, Anagnostou E, Arango C, Auzias G, Behrmann M, Busatto GF*, et al.* Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group. *Am J Psychiatry* 2018, **175**(4)**:** 359-369.

67. Chisholm K, Lin A, Abu-Akel A, Wood SJ. The association between autism and schizophrenia spectrum disorders: A review of eight alternate models of co-occurrence. *Neurosci Biobehav R* 2015, **55:** 173-183.

68. Spek AA, Wouters SGM. Autism and schizophrenia in high functioning adults: Behavioral differences and overlap. *Res Autism Spect Dis* 2010, **4**(4)**:** 709-717.

69. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J*, 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* 2015, **12**(3).

70. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME*, et al.* The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain imaging and behavior* 2014, **8**(2)**:** 153-182.

71. Trzaskowski M, Shakeshaft NG, Plomin R. Intelligence indexes generalist genes for cognitive abilities. *Intelligence* 2013, **41**(5)**:** 560-565.

72. Steinmetz H, Seitz RJ. Functional anatomy of language processing: neuroimaging and the problem of individual variability. *Neuropsychologia* 1991, **29**(12)**:** 1149-1161.

73. Fischl B, Rajendran N, Busa E, Augustinack J, Hinds O, Yeo BT*, et al.* Cortical folding patterns and predicting cytoarchitecture. *Cereb Cortex* 2008, **18**(8)**:** 1973-1980.

74. Mazoyer B, Mellet E, Perchey G, Zago L, Crivello F, Jobard G*, et al.* BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage* 2016, **124**(Pt B)**:** 1225-1231.

75. Helfinstein SM, Schonberg T, Congdon E, Karlsgodt KH, Mumford JA, Sabb FW*, et al.* Predicting risky choices from brain activity patterns. *Proc Natl Acad Sci U S A* 2014, **111**(7)**:** 2470-2475.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

76. Mahowald K, Fedorenko E. Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage* 2016, **139:** 74-93.

77. Scott TL, Gallee J, Fedorenko E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive neuroscience* 2017, **8**(3)**:** 167-176.

78. Blank I, Kanwisher N, Fedorenko E. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *J Neurophysiol* 2014, **112**(5)**:** 1105-1118.

79. Fedorenko E, Duncan J, Kanwisher N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci U S A* 2013, **110**(41)**:** 16616-16621.

80. Leech R, Kamourieh S, Beckmann CF, Sharp DJ. Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. *J Neurosci* 2011, **31**(9)**:** 3217-3224.

81. McKiernan KA, Kaufman JN, Kucera-Thompson J, Binder JR. A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. *J Cogn Neurosci* 2003, **15**(3)**:** 394-408.

82. Park DC, Polk TA, Hebrank AC, Jenkins LJ. Age differences in default mode activity on easy and difficult spatial judgment tasks. *Front Hum Neurosci* 2010, **3:** 75.

83. Zhen Z, Kong XZ, Huang L, Yang Z, Wang X, Hao X*, et al.* Quantifying the variability of scene-selective regions: Interindividual, interhemispheric, and sex differences. *Hum Brain Mapp* 2017, **38**(4)**:** 2260-2275.

84. Wang X, Zhen Z, Song Y, Huang L, Kong X, Liu J. The Hierarchical Structure of the Face Network Revealed by Its Functional Connectivity Pattern. *J Neurosci* 2016, **36**(3)**:** 890-900.

85. Zhen Z, Yang Z, Huang L, Kong XZ, Wang X, Dang X*, et al.* Quantifying interindividual variability and asymmetry of face-selective regions: a probabilistic functional atlas. *Neuroimage* 2015, **113:** 13-25.

86. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA*, et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 2012, **489**(7416)**:** 391-399.

87. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, Szafer A*, et al.* Transcriptional landscape of the prenatal human brain. *Nature* 2014, **508**(7495)**:** 199-206.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

88. Zhang Y, Chen K, Sloan SA, Bennett ML, Scholze AR, O'Keeffe S*, et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* 2014, **34**(36)**:** 11929-11947.

89. Doucet G, Naveau M, Petit L, Delcroix N, Zago L, Crivello F*, et al.* Brain activity at rest: a multiscale hierarchical functional organization. *J Neurophysiol* 2011, **105**(6)**:** 2753-2763.

90. Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML*, et al.* Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform* 2011, **5:** 13.

91. Schwarz AJ, McGonigle J. Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. *Neuroimage* 2011, **55**(3)**:** 1132-1146.

92. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015, **11**(4)**:** e1004219.

93. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG*, et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012, **491**(7422)**:** 56-65.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Fig. 1 shows a schematic of our approach for measuring the correlation between functional connectivity and gene expression profiles within a given network of brain regions. This analysis pipeline consisted of a) defining sets of cortical regions using task activation data (directly or via meta-analysis), b) estimating the resting-state functional connectivity and transcriptomic networks among these regions, and c) assessing the similarity between the functional connectivity and transcriptomic networks, along with subsequently d) estimating each gene's individual contribution (see below). The details of each dataset and procedure are described below.

Datasets

fMRI dataset acquisitions were approved by the Institutional Review Board of each site. Written informed consent was obtained when necessary from all participants, before they took part.

BIL&GIN. This dataset included 144 healthy right-handed adults (aged 27±6 years; 72 females) drawn from the larger BIL&GIN database which is roughly balanced for handedness [74]. Each participant completed three slow-event fMRI runs (gradient echo planar imaging, TR = 2.0 s, acquisition voxel size = 3.75 × 3.75 × 3.75 mm^3; 3T Philips Intera Achieva scanner) in which they were asked to complete 3 different sentence tasks including covertly producing, listening to, or reading sentences and familiar word lists as reference. Of these participants, 137 also completed rs-fMRI scans using the same imaging sequence as that used for the tasks, which lasted 8 minutes (240 volumes). Immediately prior to rs-fMRI scanning, the participants were instructed to "keep their eyes closed, to relax, to refrain from moving, to stay awake and to let their thoughts come and go". Note that the latter scanning session took place around 1 year before the task fMRI scans. For more details see further below, and [35].

NeuroSynth. Neurosynth (http://neurosynth.org) is a platform for large-scale synthesis of task fMRI data [41]. It uses text-mining techniques to detect frequently used terms as proxies for concepts of interest in the neuroimaging literature: terms that occur at a high frequency in a given study are associated with all activation coordinates in that publication, allowing for automated term-based meta-analysis. Despite the automaticity and potentially high noise resulting from the large-scale meta-analysis, this approach has been shown to be robust and meaningful (e.g., [9, 41, 44, 75]), due to the high number of studies included. We used database version 0.6 (current as of July 2018) which included 413,429 activation peaks reported in 11,406 studies (see below for the search terms employed).

EvLabN60. This dataset included statistical maps of the task fMRI contrast for passively reading sentences versus non-words from 60 participants (aged from 19 to 45; 41 females; all right-handed). This fMRI task (TR = 2.0 s, acquisition voxel size 2.1 × 2.1 × 4.0 mm^3; 3 T Siemens Trio scanner)

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was designed to localize the sentence processing network (for details, see Fedorenko et al., 2010). The sentences > non-words contrast has been previously shown to reliably activate language-sensitive regions and to be robust to the materials, task, and modality of presentation [40, 42, 76, 77]. In addition, the EvLabN60 dataset was used to define two other networks used as comparisons to sentence processing networks (below). A spatial working memory task was designed to localize the fronto-parietal multiple demand system (i.e., contrast Hard versus Easy) [78, 79] and the default mode network (i.e., contrast Easy versus Hard) [80, 81, 82]. Participants were instructed to keep track of four (Easy condition) or eight (Hard condition) sequentially presented locations in a 3×4 grid. In both conditions, participants performed a two-alternative forced-choice task at the end of each trial to indicate the set of locations they just saw. For more details, see Fedorenko et al., 2011.

GEB. GEB (http://www.brainactivityatlas.org), which is an abbreviation for "Gene-Environment-Brain-Behavior", provided an independent rs-fMRI dataset for the present study. GEB is an on-going project that focuses on linking individual differences in human brain and behaviors, to environmental and genetic factors [9, 83, 84, 85]. Rs-fMRI data from forty college students (20 females; aged = 20.3 ± 0.91 years) were included in this study. The resting-state scan lasted 8 min and consisted of 240 contiguous echo-planar-imaging (EPI) volumes (TR = 2.0 s; acquisition voxel size = 3.125 × 3.125 × 3.6 mm^3; 3 T Siemens Trio scanner). During the scan, participants were instructed to relax and remain still, with their eyes closed. The dataset has high quality in terms of minimal head motion and registration errors, and has been used in several previous studies (e.g., [44, 84]).

AHBA. AHBA (Allen Human Brain Atlas; http://www.brain-map.org) is a publicly available online resource for gene expression data. The atlas characterizes gene expression in postmortem human brain with genome-wide microarray-based data including over 20,000 genes for ~500 sampling sites distributed over the whole brain. See [86] for more details about the data collection. Normalized microarray data were used in the present study. To date (search conducted on Mar. 30, 2017), six adult donors with no history of neuropsychiatric or neurological conditions were available in the database (age 24, 31, 34, 49, 55, and 57 years; 1 female). Left hemisphere cerebral cortical data are available for all six donors whereas right-hemisphere data are available for only two of them. Detailed information on donors and analysis methods is available at www.brain-map.org. Structural brain imaging data of each donor were used to align sampling sites into standard coordinate space. We also used data from the Allen Institute's BrainSpan project (http://www.brainspan.org/), which includes human brain tissues from age 8 weeks post conception to 40 years, sampling an average 13 regions (range, 1-17) from one to three brains per time point, and measured using RNA-sequencing [87]. We used these latter data for examining the expression of specific genes of interest across human brain development.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single cell RNA-sequencing data for mouse cerebral cortex. Cell-type-specific expression levels, as indexed as fragments per kilobase of transcript, per million fragments sequenced (FPKM), were obtained to investigate the preferential expression pattern of genes of interest [88] (http://www.stanford.edu/group/barres_lab/brain_rnaseq.html). In addition, we employed a more recently available single-cell gene expression database [50] to investigate cell-type specificity.

GWAS results for ASD, Schizophrenia and intelligence. We downloaded summary statistics from the Psychiatric Genomics Consortium (http://www.med.unc.edu/pgc) for ASD with up to 7387 cases and 8567 controls (ASD GWAS 2017) [51], and schizophrenia with up to 36,989 cases and 113,075 controls (PGC-SCZ2) [32]. We also downloaded GWAS association results for intelligence, based on 78,308 individuals from the UK Biobank, CHIC consortium, and five additional cohorts (https://ctg.cncr.nl/software/summary_statistics) [52].

Defining Cortical Regions for Sentence Processing

We employed three strategies and data sources for defining the cortical sentence processing network. To refer to the regions defined under each of these three approaches, we will use the terms Supramodal Sentence Areas (SmSA), Synthesized Sentence Areas (SSA), and One-contrast Sentence Areas (OcSA). Given left-hemisphere dominance of language network (see Introduction) and limited post mortem gene expression data for the right hemisphere (see above), we focused on the left hemisphere in the present study.

SmSA. We applied the definition of left-hemispheric high-order and supramodal sentence areas provided by The FALCON atlas [35]. This atlas of language integrative and supramodal areas, involving 142 healthy rigth-handers, is based on the conjunction of activation across sentence production, listening and reading, as contrasted with activation for lists of overlearned words (again presented as either production, listening or reading tasks) in the same participants. Then, a second criterion was applied whereby leftward activation asymmetry was required during the 3 sentence minus words contrasts. See Labache et al. (2018) for a full description of this definition approach. Task fMRI data were used from the BIL&GIN dataset (see above) for this purpose. In order to obtain accurate measures of functional asymmetry, this work is based on the use of the AICHA atlas, including left-right homotopic regions of interest based on resting-state functional connectivity data [43]. In the 179 left-right pairs of homotopic regions of the AICHA atlas, fMRI signal variation and asymmetry were calculated for each task contrast and each participant. Regions with both significant activation and leftward asymmetry across the three sentence-level versus word-list task contrasts were identified. A significance threshold of Bonferroni-corrected $p < 0.05$ was applied. Thirty-two left-hemisphere regions were obtained, including 25 cortical and 7 subcortical regions [35].

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SSA were defined based on a large-scale neuroimaging meta-analysis of fMRI studies using Neurosynth (see above). A combination of terms related to sentence processing were used, including "sentence comprehension", "sentence", and "sentences" (411 studies). The resulting meta-analysis map (i.e., the likelihood map that shows there would be activation in some specific brain regions given the presence of particular terms) was used in this study to cover regions that are relevant to the network of interest. To control the false positive rate in the statistical map, a false discovery rate (FDR) threshold was used of 0.01 on a whole brain basis. As for SmSA, the AICHA atlas was used to define the areas for functional network construction. Specifically, if a region from the AICHA atlas had more than half (50%) of its voxels showing significant specificity based on the thresholded mask from the meta-analysis (FDR-corrected p <0.01), we included that region as one of the SSA.

OcSA were defined based on the probabilistic activation map of a single fMRI contrast from the EvLabN60 dataset (see above). The passively reading task has been previously shown to reliably activate language-sensitive regions and to be robust to the materials, task, and modality of presentation [40, 42, 76, 77]. The map was created by overlapping the statistical maps from all participants for the contrast of sentences versus non-words (t >2.3) onto the MNI152 template, and then dividing by the total number of participants (e.g., [83]). The value for each voxel in the obtained map indicated the probability of the voxel showing a significant contrast activation across the population. A probability threshold of 50% was applied to identify voxels showing consistent activation (t > 2.3) across subjects, and regions from the AICHA atlas with more than half of the voxels activated were included as OcSA.

Exclusion of regions: We excluded one region which was relatively small (less than 150 voxels; i.e., G_Cingulum_Post-3) and had limited gene expression data in the AHBA (fewer than 2 sampling sites), as well as subcortical areas which are known to have very different gene expression profiles to cerebral cortex and would swamp the analysis (e.g., Hippocampus), a region that had resting-state data missing in the GEB dataset (i.e., G_Paracentral_Lobule-4), and two deep regions where the gene expression data was found to diverge substantially from most cerebral cortical regions (i.e., G_ParaHippocampal-1 and G_Insula-anterior-1) (Supplemental Fig. S1). This resulted in 21 SmSA, 22 SSA, and 12 OcSA. The same criteria were also applied for defining the comparison systems below (Supplemental Table S1).

Comparison networks. The NeuroSynth term 'navigation' was used to localize cortical regions involved in spatial navigation (64 studies; search conducted on Nov. 3, 2016). A FDR threshold of 0.01 was used to control the false positive rate. Moreover, the default mode network and the fronto-parietal multiple demand networks were defined using the EvLabN60 dataset (see above), with the probabilistic activation maps of fMRI contrast Easy versus Hard, and Hard versus Easy, respectively, based on the spatial working memory task. Again, a probability threshold of 50% was applied to identify voxels showing consistent activation (t > 2.3) across subjects. Next, the AICHA atlas was

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

used to identify regions for each functional network. Our purpose was to define a similar number of top regions for each comparison network, to support similarly-powered analyses of all networks in the downstream analyses. In order to obtain comparable numbers of top regions for each network, different thresholds of overlap were applied for each network. Specifically, we found that a threshold of 1/4 defined 19 top regions of the spatial navigation network (SNN) (see below), a threshold of 1/2 defined 12 top regions of the multiple demand network (MDN), and a threshold of 3/4 defined 17 top regions of the default mode network (DMN).

Construction of the Functional Connectivity Networks

Two independent rs-fMRI datasets, i.e. BIL&GIN and GEB (above), were used for functional connectivity network construction among the sets of regions defined based on task fMRI activation.

Functional connectivity in the BIL&GIN dataset. The preprocessing of the BIL&GIN dataset was done by the Bordeaux group (MJ). Preprocessing procedures included head motion correction, registration onto the anatomical T1 image, the latter being stereotaxic registered on the MNI152 standard space. Additionally, time series for white matter and cerebrospinal fluid, the six head motion parameters, and the temporal linear trend were removed from the stereotaxic normalized rs-fMRI data using regression analysis and time series data were temporally filtered using a least squares linear-phase finite impulse response (FIR) filter design bandpass (0.01-0.1 Hz). For each participant and each region, a time series was then calculated by averaging the rs-fMRI time series of all voxels located within that region. For each individual, we computed the Pearson's correlation coefficient between the time series of each pair of cortical regions from a given task-defined set of regions. Correlation coefficients were transformed to Gaussian-distributed z scores via Fisher's transformation. For each of the 6 networks (3 sentence processing and 3 other functional networks), the functional connectivity matrix was computed by averaging data from each individual. For a full description of the processing see [89].

Functional connectivity in the GEB dataset. The preprocessing of the GEB dataset was done by the Beijing group (JL). Preprocessing procedures included head motion correction, spatial smoothing, intensity normalization, and removal of linear trend, using the FEAT preprocessing workflow implemented with Nipype [90]. A temporal band-pass filter (0.01-0.1 Hz) was applied to reduce low frequency drifts and high-frequency noise. To eliminate physiological noise, such as fluctuations caused by motion or cardiac and respiratory cycles, nuisance signals were regressed out. Nuisance regressors were averaged cerebrospinal fluid signal, averaged white matter signal, global signal averaged across the whole brain, six head realignment parameters obtained by rigid-body head motion correction, and the derivatives of each of these signals. The 4-D residual time series obtained after removing the nuisance covariates were registered to MNI152 standard space. After preprocessing, a

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

continuous time course for each region of a given task-based functional network was extracted by averaging the time courses of all voxels within that region. Temporal correlation coefficients between the extracted time course from a given regions and those from other regions were calculated to determine the strength of the connections between each pair of regions of a given functional network at rest. Correlation coefficients were transformed to Gaussian-distributed z scores via Fisher's transformation to improve normality, resulting in a symmetric Z value matrix (i.e., functional connectivity) for each task-defined system of each participant. Due to the ambiguous biological explanation of negative correlations [91], we restricted our analyses to positive edges and set negative edges to 0. We have applied the same processing procedure in several previous studies (e.g., [44, 84]). After resting-state functional connectivity networks were obtained corresponding to each task-defined set of regions, a mean functional connectivity network for each set of regions was calculated by averaging across participants, which was then used for subsequent analyses.

Construction of the Transcriptomic Networks

Transcriptomic networks were constructed based on gene expression profiles in the human brain from the AHBA (above). Specifically, we first extracted the normalized expression scores for each gene, from each sampling site and each donor. For genes with multiple microarray probes in the AHBA data, average values were calculated per gene at each sampling site and in each donor. Given that AICHA atlas regions are defined in the standard MNI space, the location of each sampling site was then translated into the standard space using the *alleninfo* (https://github.com/chrisfilo/alleninf). Gene expression data from within a given region were then averaged per gene and across donors, to obtain a single expression measure of each gene per region. We restricted our analyses to cerebral cortical regions with at least two sampling sites summed across all donors. We also restricted our analyses to the top 5% of all genes based on differential stability across donors as assessed over the entire cerebral cortex, i.e. a set of 867 genes that had differential stability greater than 0.357, as previously calculated [47]. We also repeated our analysis using a lower stability threshold (i.e., top 10%, stability larger than 0.25; $N = 1735$ genes), in order to assess the robustness of our findings with respect to this threshold. As a negative control, we additionally repeated the analyses using the lowest 5% and lowest 10% of genes as regards differential stability across donors, with which we would expect null results.

Our processing of the genetic data produced a vector of gene expression values across regions. Transcriptional Similarity (TS) between pairs of regions was then estimated by Spearman correlation, as a measure of 'transcriptomic connectivity'.

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similarity between the Functional and Transcriptomic Networks

Network similarity analysis. To examine the similarity between functional connectivity networks and transcriptomic networks, correlation analyses were performed. Specifically, a vector was extracted from the upper triangle of the connectivity matrix of each network, and Spearman correlation was then calculated between the vectors of the functional connectivity networks and their corresponding transcriptomic networks. To rule out potential influence of spatial proximity, we also calculated the correlations between residuals of the two measures after controlling for the spatial distance between centers of regions (i.e., the Euclidean distance of MNI coordinates, available within the AICHA atlas [43]) using a regression approach.

Gene contribution index (GCI). In addition to the overall correlations between functional connectivity and corresponding transcriptomic networks, we formulated a novel index, the gene contribution index (GCI), for estimating the contribution of each individual gene to an observed overall correlation. GCI was defined as the difference in the overall correlation before and after removing that gene at the step of transcriptomic network construction, i.e. based on a 'leave-one-out' approach.

Identification of 'consensus genes' correlated with the sentence processing network. We identified a set of 'consensus genes' (N = 41) which had positive CGI scores in all six analyses of the sentence processing network, i.e. the three definition strategies SmSA, SSA, and OcSA, by the two independent rs-fMRI datasets BIL&GIN and GEB. For each of the three independent comparison networks (spatial navigation, fronto-parietal multiple demand, and default mode networks) we also identified the same number of top genes (i.e., 41) showing the highest GCIs, based on the averaged score of each gene in the two analyses relevant to each of those networks, i.e., with the two rs-fMRI datasets.

Follow-up Analyses of the Consensus Genes

We further investigated the consensus genes correlated with the sentence processing network by use of literature searches and bioinformatics tools:

Gene ontology. The gene ontology provides a classification scheme for genes based on what is known with respect to their molecular functions, the biological processes that they are involved in, or cellular components that they encode (http://www.geneontology.org/). Gene ontology analyses were performed with the Bioconductor package gProfileR (https://biit.cs.ut.ee/gprofiler/), using ontologies from Ensembl release 91. Gene sets containing between 25-1000 genes were included. All known genes were used for determining the statistical domain size in the analysis. The default g:SCS method in the tool was used for multiple testing correction (corrected $p < 0.05$).

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Expression across brain development. We queried the Allen Institute's BrainSpan project data (see Datasets above) for the consensus genes, in relation to their developmental changes in expression.

Gene expression specificity. For the 41 consensus genes correlated with the sentence processing network, we contrasted their expression levels between those cortical regions that were assigned to the sentence processing network under any definition (i.e. all SmSA, SSA and OcSA) (N = 31), versus all regions outside the system that were included among the SNN, MDN or DMN (N = 34). More information about the lists of regions can be found in Supplemental Table S1. The independent samples t-test (equal variances not assumed) was used to contrast expression levels. Since multiple comparisons were performed across 41 genes, a significance threshold of FDR-corrected p value of 0.05 was applied (Benjamini-Hochberg FDR).

In terms of cell-type specificity, the expression levels of each of the 41 consensus genes correlated with the sentence processing network, were queried in a published dataset based on mouse cortical data as indexed as reads per kilobase of exon per million reads mapped (FPKM) [88]. We also queried another single-cell gene expression database [50] to investigate the cell-type specificity of genes.

Enrichment Analysis using GWAS Summary Statistics for ASD, Schizophrenia and Intelligence

We tested the hypothesis that the consensus set of sentence processing network genes was enriched for association signals with ASD, schizophrenia or intelligence. (There were no GWAS results based on more than 10,000 subjects yet available for reading/language measures in the general population, or disorders such as dyslexia or language impairment which involve linguistic deficits.) Specifically, we ran gene set analyses using the MAGMA software (Version 1.06; http://ctg.cncr.nl/software/magma) [92]. MAGMA was run with default settings and no gene extension window. Briefly, gene-based association scores were derived using the SNP-wise mean model, which considers the sum of -log(p-values) as derived from GWAS analysis, for single nucleotide polymorphisms (SNPs) located within the transcribed region of a given gene (using NCBI 37.3 gene locations). MAGMA accounts for gene-size, number of SNPs in a gene, and linkage disequilibrium (LD) between SNPs when estimating gene-based association scores. LD between SNPS was based on the 1000Genomes phase 3 European ancestry samples [93]. In this analysis, the score for a given gene therefore indicates how strongly genetic variation within, or in linkage disequilibrium with, that gene is associated with the trait of interest (i.e., ASD, schizophrenia, or intelligence). These GWAS-based gene scores were subsequently used to compute gene set enrichment within the 41 consensus genes associated with the sentence processing network. The enrichment analysis tests whether the genes in a given set have, on average, higher GWAS-based gene scores than the other genes in the genome. No cutoff was made on the GWAS-based gene scores, so that for all genes the degree of association with the trait of interest was taken into account. Similarly, we conducted the gene set analysis with top

bioRxiv preprint doi: https://doi.org/10.1101/439984; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes for the other comparison functional networks, to compare with results for the sentence processing network.

Acknowledgements

We thank the NeuroSynth, Allen Human Brain Atlas, and the Psychiatric Genomics Consortium for data sharing. This study was funded by the Max Planck Society (Germany).

Author Contributions

Conceived and designed the experiments: XZK CF. Performed the experiments: XZK NTM MJ EF JL. Analyzed the data: XZK MJ. Contributed data/materials/analysis tools: XZK NTM MJ EF JL. Contributed to the writing of the manuscript: XZK NTM MJ EF JL SEF CF.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

BIL&GIN: NTM and MJ; EvLabN60: EF; GEB: XZK and JL; NeuroSynth: XZK via http://neurosynth.org/; AHBA: XZK via https://www.brain-map.org/

Data Availability

The authors declare that the data supporting the findings of this study are available within the article and its supplementary files. All relevant data are available from the authors upon request.