Linking central gene expression patterns and mental states using transcriptomics and large-scale meta-analysis of fMRI data: A tutorial and example using the oxytocin signaling pathway

Jaroslav Rokicki ^{a,b,‡}, Daniel S. Quintana ^{a,‡*} and Lars T. Westlye ^{a,b}

^a Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, University of Oslo, and Oslo University Hospital, Oslo, Norway

^b Department of Psychology, University of Oslo, Oslo, Norway

[‡] These authors contributed equally

* **Corresponding author:** Daniel S. Quintana, Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, University of Oslo, and Oslo University Hospital, Oslo, Norway. Email: daniel.quintana@medisin.uio.no.

Running Head: Linking central gene expression patterns and mental states

Keywords: social behavior, social cognition, neuroendocrinology, oxytocin, gene expression, oxytocin receptor

Abstract

The measurement of gene expression levels in the human brain can help accelerate our understanding of complex mental states and psychiatric illnesses. Mental states are typically associated with whole-brain networks, however, gene expression levels from post-mortem brain samples have traditionally been measured in a limited number of brain regions due to resource limitations. The recent availability of whole-brain gene expression data from the Allen Human Brain Atlas (AHBA) provides the opportunity to generate gene expression patterns for over 20,000 genes. By linking these expression patterns with brain activity patterns that are associated with specific mental states, researchers can better understand which genes may support given mental states, via forward inference. Conversely, reverse inference can also be used to determine which mental state activation patterns are most strongly associated with a given gene expression map. This chapter provides a step-by-step guide on how to use the AHBA in conjunction with the NeuroSynth fMRI meta-analysis tool to identify the mental state correlates of specific gene expression patterns, using genes from oxytocin signaling pathway as an example. We also demonstrate how to perform an out-of-sample validation and assess the specificity of results for genes of interest.

1. Introduction

The development and maintenance of the structural and functional properties of the human brain that are required to support complex mental functions is partly controlled by a dynamic and distributed pattern of gene activity across the brain. This pattern of gene regulation can be indexed by measuring transcriptional activity. Given that several neural networks underlying complex mental states have been well-characterized [1], identifying gene expression via mRNA levels across the brain can be used to help unravel the molecular basis of brain function and cognition. The emergence of imaging transcriptomics has facilitated the linking of the brain’s transcriptome with brain imaging measures. Gene expression levels have historically been measured in a handful of brain regions [e.g., 2], due to resource and technical limitations. While such analyses have been instructive, incomplete maps cannot provide the level of detail required to link gene expression to complex whole-brain neural networks.

The Allen Human Brain Atlas (AHBA; <http://human.brainmap.org/>) provides gene expression data for over 20,000 genes from almost 4,000 anatomical regions across almost the entire human brain mapped onto stereotaxic space [3]. Analysis of the AHBA has revealed that genes with considerable biological significance, such as those associated with calcium signaling (e.g., *NRGN*, *CAMK2A*, *CAMKV*), demonstrate high conservation between individuals [4]. Moreover, the expression patterns of genes with high biological relevance are also highly correlated with functional connectivity patterns [4]. Given the high spatial resolution of the AHBA, expression patterns have

been linked to several brain imaging modalities, such as anatomic connectivity [5], cortical thickness [6], and positron emission tomography [7].

The development of large-scale meta-analysis tools for functional neuroimaging data, such as NeuroSynth [8], provides a means to synthesize a large body of neuroimaging findings and identify summary brain activity patterns that are associated with specific cognitive and mental states. While this approach can be used for forward inference, in which a summary brain activation map can be generated for a given mental state, reverse inference can also be applied, in which mental states can be identified that are associated with a given brain map. The spatial resolution of the AHBA provides the opportunity to assess which mental and mental states are associated with a given gene expression pattern. In this chapter, we demonstrate how to link gene expression patterns from the AHBA to mental states. Identifying gene expression patterns that are associated with given mental states can help unravel the molecular basis of complex human cognition.

To demonstrate this method, we will use our work on the oxytocin signaling pathway as an example [9]. The neuropeptide oxytocin is well-known for its effects on parturition and the milk let-down reflex. However, more recently, research has begun investigating oxytocin's role in social behavior and cognition [10, 11], with potential theoretical, mechanistic and therapeutic applications for psychiatric disorders characterized by social dysfunction. Given that oxytocin receptor expression patterns in the mammalian brain are associated with behavior [12, 13], identifying the location

of oxytocin receptor genes can help better identify the role of this evolutionarily ancient signaling system.

2. Materials and Methods

Five main steps are required to calculate gene expression brain pattern maps, the biological relevance of these patterns, and the associations of these pattern maps with mental state brain activations: i) Gene expression data processing, ii) Statistical inference, (iii) Out-of-sample validation iv) Weighted gene-expression analysis, and v) Calculating mental state correlates using NeuroSynth. Each of these steps are will now be addressed in turn.

2.1. Gene expression data processing

The first step in this analysis pipeline is to download and prepare the AHBA dataset. This dataset contains central gene expression data from six donors. Three donors were Caucasian males, one donor was a Hispanic female, and two donors were African-American males. The mean donor age was 42.5 (S.D. = 11.2) years and brain tissue was collected 22 hours after demise, on average (S.D. = 4.5). The following steps are required for processing the AHBA data:

- Download the complete normalized microarray datasets from <http://human.brain-map.org/>. These datasets are divided per donor with the

following files: a “MicroarrayExpression.csv” file with gene expression values, a “Probes.csv” file with detailed probe descriptions, and a “SampleAnnot.csv” file with brain region descriptions.

- Due to inconsistencies in the brain regions sampled across donors (ranging from 363 to 473 regions per left hemisphere), a few steps are required to interpolate missing locations to generate voxel-by-voxel gene expression maps. MATLAB scripts to implement interpolation can be accessed at <https://osf.io/kgtqd/>, but we also describe this method as follows. First, map samples onto a brain scan in the native space of each subject (1 mm isotropic space). Second, fill the whole left hemisphere with valid expression values by labelling brain borders with the sample expression value that has the closest physical distance to a given border point (Fig. 1A). Third, linearly interpolate brain voxels within the brain with missing expression values. This is accomplished by dividing the space between scattered points into simplices based on Delaunay triangulation (Fig. 1B), then linearly interpolating each simplex with values to yield a completed map (Fig. 1C).
- To compare donors, brains can be registered with voxel-by-voxel expression maps mapped to MNI space using ANTs (<http://stnava.github.io/ANTs/>) or FNIRT non-linear registration (<https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FNIRT>).
- For many genes in the AHBA dataset, more than one probe is available that represent different fragments of the same gene. In such cases, differential stability

values can be used to select the probe with the most consistent expression pattern from donor-to-donor. Moreover, differential stability can be used to examine the *relative* level of expression stability across all protein coding genes. Genes with higher differential stability tend to have strong biological relevance [4]. Moreover, the patterns of genes with high differential stability in the AHBA dataset are highly conserved, thus can be better extrapolated to the human population in general. Genes with low differential stability in the AHBA dataset are indicative of high interindividual differences in gene expression patterns, which are less generalisable. Differential stability is calculated using an approach described by Hawrylytz and colleagues [4], which quantifies pattern reproducibility across different subjects. Three steps are required to calculate differential stability. First, calculate Spearman’s correlation coefficient for each possible combination of 6 brains (15 pairs) using MNI registered left hemisphere of voxel-by-voxel brain maps created in previous steps. Second, average differential stability over 15 possible combinations between subjects. Differential stability can theoretically range from 1 (complete agreement) to 0 (no agreement). In the AHBA sample, the most stable gene is *TNNT2*, with a differential stability value of 0.81. Third, segment the expression levels for the probes of interest using an atlas of choice (e.g., Automated Anatomical atlas) once the probe with the highest differential stability is identified.

2.2. Statistical inference

Using a brain atlas facilitates visualisation of expression levels across the brain (Fig. 2). However, statistical inference can also be applied to assess whether any regions have statistically significantly higher or lower expression compared to average expression across the brain. Given the number of tests that are being conducted when using brain imaging atlases, these tests should be corrected for multiple tests to reduce the risk of false positives. An example analysis script for R, which includes code to complete the steps below, is located at <https://osf.io/vu2d6>.

- The atlas of interest must be in the same space as the expression data (e.g., MNI space, 1mm isotropic voxels). To accomplish this, first calculate the median over each region of interest (ROI) within selected atlas. That is, if the selected atlas has 118 regions, filter out ones on the left hemisphere first (in the case of AAL atlas it leaves us with 54 regions), and subsequently apply this to each of 6 donors. Second, calculate the median or average, and the standard deviation for 6 donors. Third, calculate the median or average within all ROIs in the given atlas.
- Conduct one-sample t -tests to assess which of the left hemisphere regions (as calculated in the previous step) from the six donor samples expressed mRNA to a significantly greater or lesser degree compared to average mRNA expression across the brain.

- Correct for multiple tests (number of ROIs in left hemisphere) using a false discovery rate (FDR) threshold.
- Calculate Cohen's d values for one-sample t -tests to yield measures of effect size.

2.3. Out-of-sample validation

The Genotype-Tissue Expression (GTEx) database can be used for out-of-sample validation of results from the AHBA. The GTEx database has less spatial accuracy than the AHBA, but it has a much higher number of donors (mean sample size for mRNA expression across brain regions = 131.7, range = 88 to 173). Out-of-sample validation can be conducted using the following steps:

- Download the GTEx database (<https://gtexportal.org/home/dataset>)
- Extract medians for brain region expression in Log₁₀ transcripts per million for 10 distinct brain regions: amygdala, anterior cingulate cortex (BA24), caudate (basal ganglia), cerebellar hemisphere, cerebellum, cortex, frontal cortex (BA9), hippocampus, hypothalamus, nucleus accumbens (basal ganglia), putamen (basal ganglia), and substantia nigra.
- Extract medians from the AHBA dataset from the raw data by reproducing an atlas with same anatomical regions as ones available in GTEx database.
 - Use the AAL atlas for amygdala, caudate, cerebellum (by merging all cerebellum sub-regions), hippocampus and putamen regions

- Use the Talairach atlas for the Anterior cingulate cortex (BA24), frontal cortex (BA9), hypothalamus and substantia nigra.
 - Use the Hammersmith atlas for the Nucleus Accumbens
- With the new atlas, calculate average expression values for each ROI (as described above) by substituting the AAL atlas with the newly created one.
- Calculate Spearman's rank correlation coefficients to assess the relationship between oxytocin pathway of interest expression profiles from the two databases and overall co-expression patterns in these 10 regions.
- Since the AHBA database contains microarray expression and GTEx contains RNA-sequencing data, one cannot compare these two sets directly. Alternatively, construct a co-expression matrix separately for each source and then investigate and compare the interactions between genes.
- Select list of genes you want to compare. For instance, in Quintana et al [9], a matrix was constructed using selected set of genes: oxytocin pathway set (*OXT*, *CD38*, *OXT*); dopaminergic set (*DRD1*, *DRD2*, *DRD3*, *DRD4*, *DRD5*, *COMT*, and *DAT1*); muscarinic acetylcholine set (*CRHM1*, *CRHM2*, *CRHM3*, *CRHM4*, and *CRHM5*); vasopressin set (*AVPR1A*, *AVPR1B*); *opioid set* (*OPRM1*, *OPRD1*, *OPRK1*).
- Use complete linkage clustering [14] to identify clustering groups in both datasets for comparison.

- To get functional description of identified clusters, upload gene lists to the FUMA tool (<https://fuma.ctglab.nl/>).

2.4. Weighted gene expression analysis

The WGCNA R package [15] can be used to reveal gene co-expression networks within larger gene sets (e.g., the oxytocin signaling pathway), as per the following steps:

- Input the expression data used above via the AAL atlas for the left hemisphere. Thus, each donor will be represented by expression values in 54 regions. Another option is to use raw expression values, but then different donors will be represented by a variable number of expression measurements.
- While constructing a weighted gene network, it is important to ensure that the network constitutes a biologically meaningful model, or in other words, that it follows a power law. To do that, one needs to choose a proper soft thresholding power parameter, which is used to raise the co-expression to calculate adjacency. This can be achieved using the WGCNA function ‘pickSoftThreshold’. Choose power at which scale free topology model fit reaches 0.80. In Quintana et al [9], we selected a power of 10.
- From the expression data, calculate the adjacency matrix using the function ‘adjacency’ using the power value used in the previous step.

- From the adjacency matrix, calculate a topological overlap matrix using the function ‘TOMsimilarity’ and subtract it from 1 to get the distance matrix.
- To identify clusters of gene modules with tight interconnections, use ‘hclust’ function on the distance matrix.
- To get functional description of clusters upload them to FUMA tool (<https://fuma.ctglab.nl/>).

2.5. Mental state correlates of central gene expression

The NeuroSynth framework facilitates the meta-analysis of brain activations maps for specific mental states by drawing data from a corpus of over 14,000 fMRI studies [8].

As well as creating synthesized maps using forward inference, it is also possible to decode the mental states associated with brain activation maps using reverse inference.

The following steps describe how to use this reverse inference approach to identify which mental state brain activation maps are most highly correlated with gene expression maps of interest.

- To calculate mental state correlates of gene expression data, use the left hemisphere voxel maps generated using the instructions above.
- Download the NeuroSynth code (<https://github.com/neurosynth/neurosynth>).

The NeuroSynth package uses a Pearson correlation coefficient by default, but

the package can be modified to use a Spearman correlation coefficient, as done in Quintana et al [9].

- Convert voxel-by-voxel expression maps to voxel-by-voxel Z score maps by subtracting the left hemisphere’s global mean and dividing by its standard deviation for each voxel within that map.
- Average voxel-by-voxel expression maps over the six AHBA donors, so that one gene is represented by a single gene expression map.
- Decode association Z maps from each mental state (e.g. “Motivation”). Each map reflects Z-scores of the association between the presence of activation and the presence of a feature. Use a “mask” option to limit results in the association maps to the left hemisphere only.
- Calculate Spearman’s correlations between the Z-score association map for a mental state, and Z-score map representing the averaged gene expression among the six donors.
- Rank the resulting correlations in descending order for each gene and select the top correlated terms per gene of interest.
- While these correlations between genes of interest and mental states are informative, the *relative* strength of these associations are unclear with this information alone. That is, while there may be a strong absolute relationship between the expression pattern of a given gene and a given mental state, perhaps thousands of other genes demonstrate stronger relationships. Thus, additional

steps are required to assess the relative strength of associations. First, calculate Spearman's correlation between association map for each of mental top 5 terms and voxel-by-voxel gene expression maps of the remainder 20,737 genes. Second, with the full distribution of Spearman's correlation coefficients for each mental state, calculate the rank of the gene of interest. This can also be visualised by visually identifying the gene of interest on a distribution of all protein-coding genes (e.g., Fig. 3).

Altogether, these five steps provide the means to better understand the genetic associations of a wide range of mental states, which can be validated using out-of-sample data.

Acknowledgements

This research was supported by an Excellence Grant from the Novo Nordisk Foundation (NNF16OC0019856), the Research Council of Norway (249795, 298646), ERA-Net Cofund through ERA PerMed (IMPLEMENT project), and the European Research Council under the European Union's Horizon 2020 research and innovation program (ERC Starting Grant 802998).

References

1. Zimmermann J, Griffiths JD, McIntosh AR (2018) Unique mapping of structural and functional connectivity on cognition. *J Neurosci* 38:9658–9667
2. Loup F, Tribollet E, Dubois-Dauphin M, Dreifuss JJ (1991) Localization of high-affinity binding sites for oxytocin and vasopressin in the human brain. An autoradiographic study. *Brain Res* 555:220–232
3. Hawrylycz M, Lein ES, Guillozet-Bongaarts AL, et al (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489:391
4. Hawrylycz M, Miller JA, Menon V, et al (2015) Canonical genetic signatures of the adult human brain. *Nat Neurosci* 18:1832
5. Goel P, Kuceyeski A, LoCastro E, Raj A (2014) Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. *Hum Brain Mapp* 35:4204–4218. <https://doi.org/10.1002/hbm.22471>
6. Shin J, French L, Xu T, et al (2018) Cell-Specific Gene-Expression Profiles and Cortical Thickness in the Human Brain. *Cereb Cortex* 28:3267–3277. <https://doi.org/10.1093/cercor/bhx197>
7. Rizzo G, Veronese M, Heckemann RA, et al (2014) The Predictive Power of Brain mRNA Mappings for in vivo Protein Density: A Positron Emission Tomography Correlation Study. *J Cereb Blood Flow Metab.* <https://doi.org/10.1038/jcbfm.2014.21>
8. Yarkoni T, Poldrack RA, Nichols TE, et al (2011) Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods* 8:665
9. Quintana DS, Rokicki J, van der Meer D, et al (2019) Oxytocin pathway gene networks in the human brain. *Nat Commun* 10:668. <https://doi.org/10.1038/s41467-019-08503-8>
10. Liu Y, Li S, Lin W, et al (2019) Oxytocin modulates social value representations in the amygdala. *Nat Neurosci* 22:633–641. <https://doi.org/10.1038/s41593-019-0351-1>
11. Quintana DS, Westlye LT, Hope S, et al (2017) Dose-dependent social-cognitive effects of intranasal oxytocin delivered with novel Breath Powered device in adults with autism spectrum disorder: a randomized placebo-controlled double-blind crossover trial. *Transl Psychiatry* 7:e1136

12. Keebaugh AC, Barrett CE, Laprairie JL, et al (2015) RNAi knockdown of oxytocin receptor in the nucleus accumbens inhibits social attachment and parental care in monogamous female prairie voles. *Soc Neurosci* 10:561–570. <https://doi.org/10.1080/17470919.2015.1040893>
13. Keebaugh AC, Young LJ (2011) Increasing oxytocin receptor expression in the nucleus accumbens of pre-pubertal female prairie voles enhances alloparental responsiveness and partner preference formation as adults. *Horm Behav* 60:498–504
14. Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. *WIREs Data Min Knowl Discov* 2:86–97. <https://doi.org/10.1002/widm.53>
15. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559

Figure captions

Figure 1. Creation of expression maps. First, brain boundaries were labelled (a).

Spaces were then divided into simplices, which were triangles in two-dimensional

cases (note that not all possible triangles are marked in the image) (b). We then

perform performed linear interpolation within each simplex (c). Figure adapted from

Quintana et al., [9] (CC BY 4.0 license)

Figure 2. Oxytocin receptor (OXTR) gene expression across 54 brain regions. Each point represents mean expression from six donors with standard errors for a given brain region for *OXTR* with the AHBA AAL atlas. The bolded dashed lines represent mean expression across the all regions with 1 standard deviation ($\pm$) also shown. Stars denote statistically significant increases or decreases from the mean ($*p < 0.05$, FDR corrected for 54 tests). Figure adapted from Quintana et al., [9] (CC BY 4.0 license).

Figure 3. Mental state correlates of central oxytocin pathway gene expression patterns. The distribution of Spearman's correlations between each protein coding gene map ($n = 20,737$) and selected mental state maps was generated, with absolute ranking of three genes of interest indicated. Figure adapted from Quintana et al., [9] (CC BY 4.0 license).