Hierarchical clustering of the human cerebral cortex based on comprehensive transcriptome data

Gryglewski Gregor, Murgas Matej, Lanzenberger Rupert

Department of Psychiatry and Psychotherapy, Medical University of Vienna, Austria

Abstract

The availability of extensive transcriptome data of the human brain has greatly expanded the possibilities to study the molecular differentiation of brain regions. A previously proposed method enabled the spatially comprehensive prediction of gene expression in the brain based on discrete microarray data. The resulting data was employed in the current work in order to derive a parcellation of the human cerebral cortex. To this end, the transcriptome dataset comprising normalized expression of 18,686 genes was used for agglomerative hierarchical clustering with Pearson correlation distance and average linkage. The optimal number of clusters indicated by the Bayesian Information Criterion was k=33. The transcriptome based parcellation was able to reproduce several well-established boundaries between cortical regions, such as primary sensory and motor areas, while revealing novel insights into their hierarchical organization.

Introduction

The evolution of brain functions which subserve several uniquely human behaviors was paralleled by the specialization of the cerebral cortex.[1] Accordingly, the parcellation of the cortex was among the earliest pursuits in modern neuroscience. It continues to advance our understanding by providing definitions of discrete areas and networks which facilitate specific scrutiny, interpretation of results and communication between researchers.[2]

This work utilizes human brain transcriptome data provided in the scope of the Allen Human Brain Atlas (AHBA) project.[3] Previously, a regression model which exploits the spatial dependence of gene expression[4] was used to generate continuous and unbiased predictions of messenger ribonucleic acid (mRNA) expression in the entire cortex from microarray data of samples collected from six adult brain donors.[5] In the current work, the resulting spatially comprehensive mRNA expression maps of 18,686 genes were used to parcellate the cortex based on transcriptomic profiles using hierarchical clustering.

This preprint version outlines the methods and rudimentary results in order to facilitate scientific discourse on the results.

Methods

Estimation of cortical gene expression

Human brain transcriptome data provided by the Allen Institute (Allen Human Brain Atlas, http://human.brain-map.org/) was downloaded and processed as described previously[5] to generate comprehensive maps of mRNA expression in the cortex which are openly available at our homepage (http://www.meduniwien.ac.at/neuroimaging/mRNA.html). In short, microarray probes with signal not significantly different from background in a minimum of 1% of samples and probes without an association with an Entrez ID were excluded, such that 40,205 probes mapped to 18,686 genes were retained. Probes associated with the same gene were averaged and individual probes were removed step-wise if their exclusion resulted in a higher spatial dependence of gene expression. For each gene, expression intensity in the six donor brains was set to an equal mean across brain regions to minimize the influence of inter-individual variation. Cortical surface reconstruction was performed using the recon-all pipeline in FreeSurfer 5.1 (Harvard Medical School, Boston, USA; http://surfer.nmr.mgh.harvard.edu/) and microarray samples from all donors were mapped to their respective nearest vertices in fsaverage space. Variogram models were fitted and Gaussian process regression (ordinary Kriging) was performed to predict mRNA expression for all surface vertices using the gstat 1.1-5 package[6] in R. This was performed for the left hemisphere, as data for the right hemisphere was not available for 4 out of 6 brains. To reduce noise, observed expression intensities of vertices associated with microarray samples were replaced by the average of the estimated expression intensity of the directly adjacent vertices. Furthermore, smoothing was applied to the mRNA expression maps with an isotropic Gaussian kernel and full width at half maximum (FWHM) of 3,4 mm which corresponds to the average Euclidean distance of vertices separated by four edges in fsaverage space. Expression maps were normalized by subtraction of the mean expression intensity of each gene across the cortex.

Parcellation of the cortex

18,686 gene expression maps were used as input for agglomerative hierarchical clustering in Matlab R2014a (https://www.mathworks.com/). Hierarchical clustering was chosen in order to reveal the hierarchical differentiation of the cerebral cortex. The distance between two vertices of the cortex was calculated from Pearson correlation coefficients (r) using the formula $d(x,y)=1/2(1-r_{x,y})$, which provides high accuracy at an acceptable computational cost.[7] A correlation based distance was used because the contrast between nearest and furthest neighbors decreases with an increasing number of dimensions for Minkowski measures, such as Euclidean or Manhattan distances, which are otherwise frequently used in clustering applications.[8] In order to select the optimal linkage method (average, complete or median linkage), robustness analysis was performed for each linkage method using data from the left cortical hemisphere. Further, the effect of z-scoring gene expression data was assessed, which was predicted to reduce robustness due to the assignment of more weight to genes with a low spatial variance in expression. Robustness analysis was performed using bootstrapping by creating 1,000 random subsets of 6,229 genes, i.e. one third of the entire transcriptome dataset, and performing clustering analysis on each of these subsets. Subsequently, the robustness of clustering results was determined by calculating the average of the Adjusted Rand Index (ARI)[9] obtained for each pair of clustering results obtained from the 1,000 subsets and for each number of clusters from k=5 to k=100. The ARI is a measure ranging between 0 and 1 which reflects the concordance between different partitions. In order to achieve computational feasibility, robustness analysis was performed on a downsampled representation of the cerebral cortex (fsaverage5) comprising 10,242 vertices.

Results

Effect of linkage and standardization method

Based on ARI, highest robustness was observed for average linkage across all cluster numbers analyzed, which is in agreement with a previous analysis of the optimal choice of linkage method for gene expression data.[7] Z-scoring reduced robustness of clustering results irrespective of linkage method. On this basis, agglomerative hierarchical clustering with Pearson correlation distance, average linkage and without z-scoring was used as the final clustering algorithm.

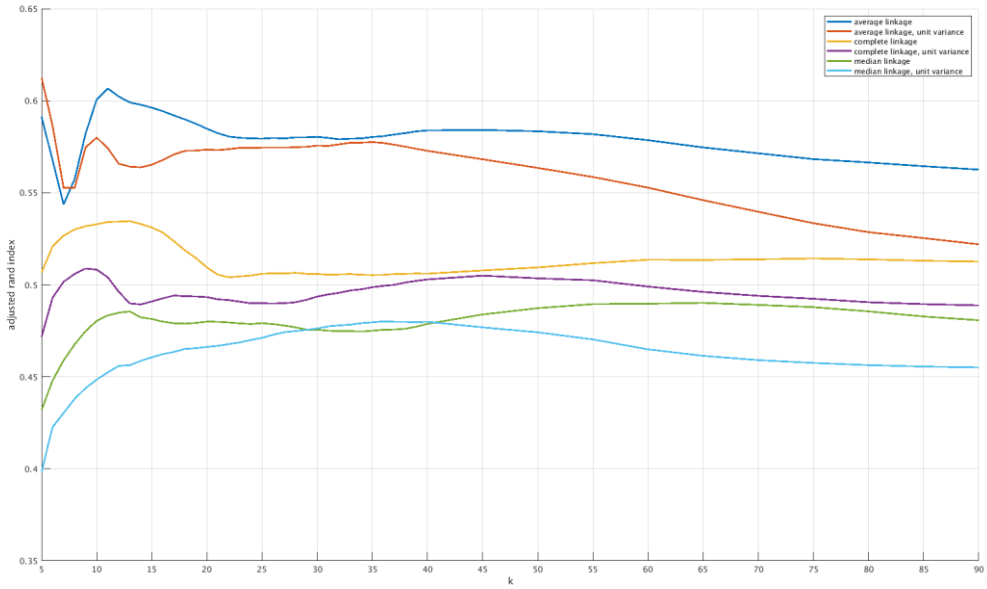


Figure 1: Effect of linkage method and standardization on robustness of clustering results.

For each linkage method, average Adjusted Rand Index (ARI) calculated for clustering results based on 1,000 random subsets comprising one third of the brain transcriptome dataset is plotted as a function of the number of clusters. This was performed for standardized (z-scored) and non-standardized (each gene's expression set to zero mean) gene expression data.

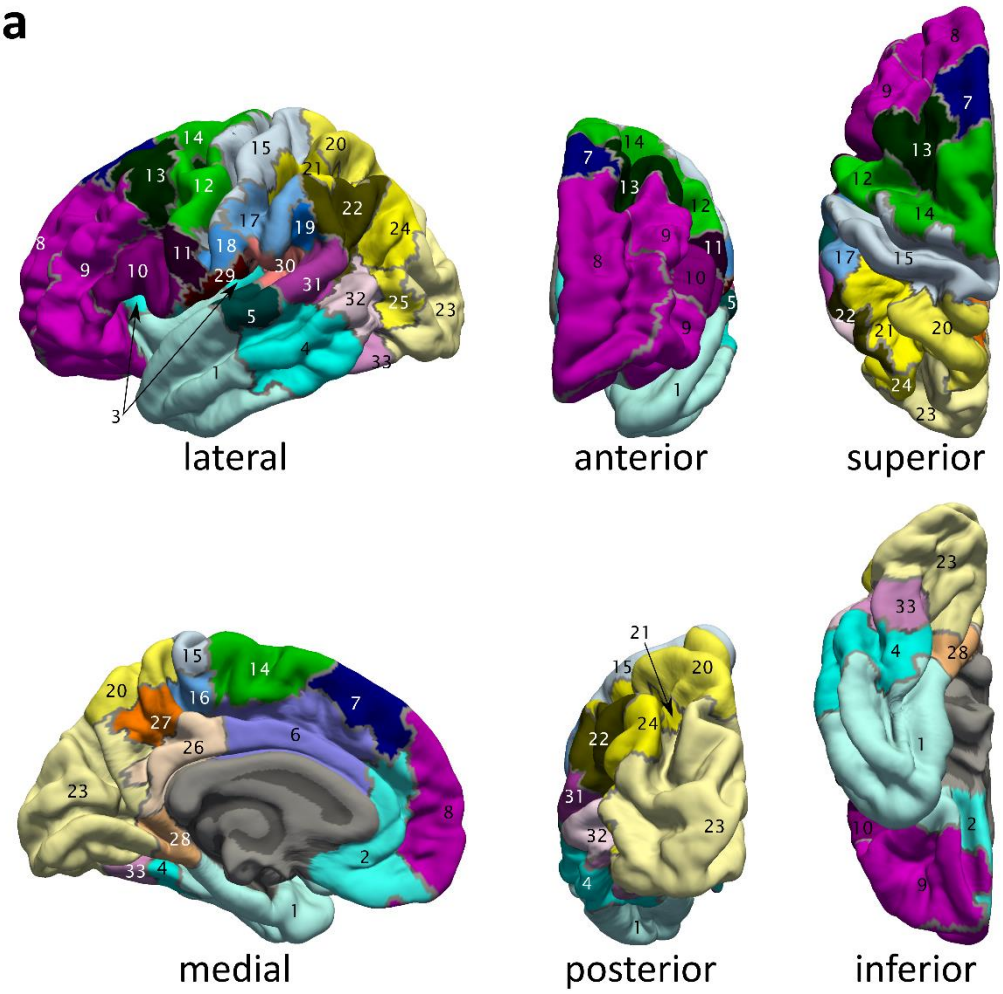
Hierarchical clustering results

The parcellation and the dendrogram created by hierarchical clustering are shown in Figure 2. The displayed parcellation at k=33 corresponds to the optimal solution indicated by the Bayesian information criterion (BIC).

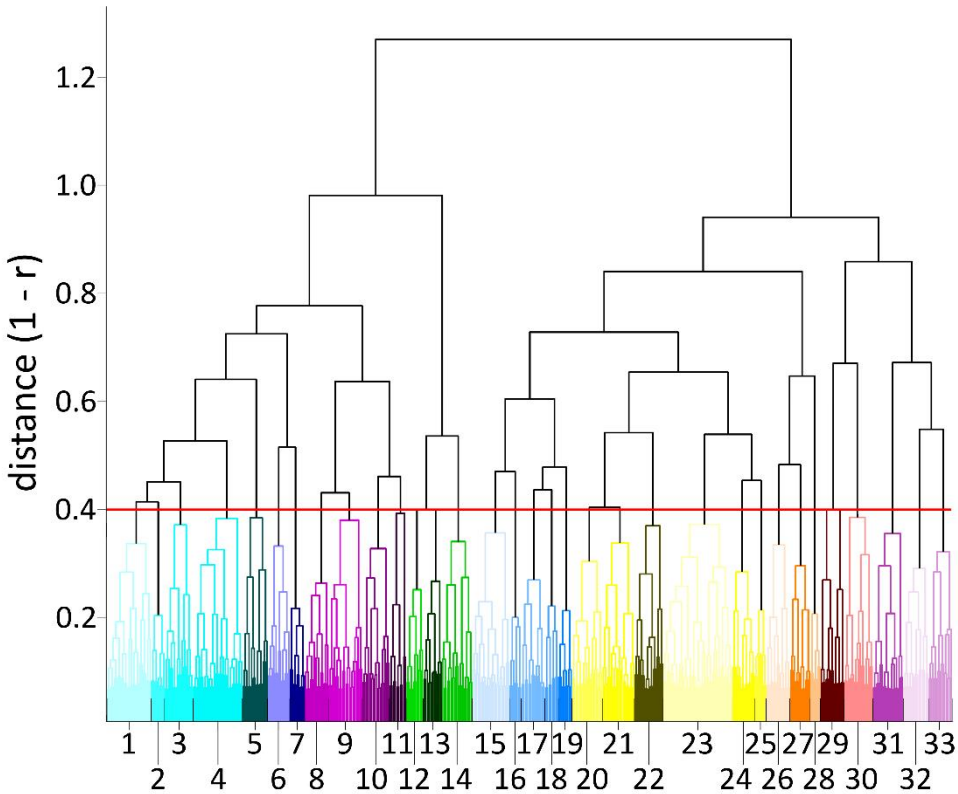
Figure 2 Transcriptome based parcellation of the human cerebral cortex.

a) The parcellation is displayed superimposed on the pial surface of the left cortical hemisphere of the average subject. **b**) The dendrogram created by the hierarchical clustering algorithm is shown with correlation distance calculated from Pearson's r plotted on the y-axis. The Bayesian Information Criterion indicated 33 as the optimal number of clusters (red line). For display purposes, nine basic colors were used. Further subdivisions down to the solution of 33 clusters are visualized by differences in brightness, darker clusters being more distant with respect to mRNA expression from the center of their respective higher-order cluster indicated by color assignment.

a


lateral
anterior
superior
medial
posterior
inferior

b


distance (1 - r)
1.2
1.0
0.8
0.6
0.4
0.2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33

Discussion

The transcriptome based parcellation was able to reproduce several well-established boundaries between cortical regions, such as primary sensory and motor areas, while revealing novel insights in their hierarchical organization.

Detailed results and their discussion will be made available in the near future. Before publication of the final paper, further data can be obtained through correspondence.

Competing interests

With relevance to this work there is no conflict of interest to declare. R. Lanzenberger received travel grants and/or conference speaker honoraria within the last three years from Bruker BioSpin MR, Heel, and support from Siemens Healthcare regarding clinical research using PET/MR. He is shareholder of BM Health GmbH since 2019.

References

1. Geschwind, D. H. & Rakic, P. Cortical evolution: Judge the brain by its cover. *Neuron* **80**, 633–647 (2013).
2. Eickhoff, S. B., Constable, R. T. & Yeo, B. T. T. Topographic organization of the cerebral cortex and brain cartography. *Neuroimage* (2018). doi:10.1016/j.neuroimage.2017.02.018
3. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
4. Romero-Garcia, R. *et al.* Structural covariance networks are coupled to expression of genes enriched in supragranular layers of the human cortex. *Neuroimage* (2018). doi:10.1016/j.neuroimage.2017.12.060
5. Gryglewski, G. *et al.* Spatial analysis and high resolution mapping of the human whole-brain transcriptome for integrative analysis in neuroimaging. *Neuroimage* **176**, 259–267 (2018).
6. Pebesma, E. J. Multivariable geostatistics in S: the gstat package. *Comput. Geosci.* **30**, 683–691 (2004).
7. Jaskowiak, P. A., Campello, R. J. G. B. & Costa, I. G. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics* (2014). doi:10.1186/1471-2105-15-S2-S2
8. Pestov, V. On the geometry of similarity search: dimensionality curse and concentration of measure. *Inf. Process. Lett.* **73**, 47–51 (2000).
9. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).