Human Brain Mapping: A Systematic Comparison of Parcellation Methods for the Human Cerebral Cortex

Salim Arslan, Sofia Ira Ktena, Antonios Makropoulos, Emma C. Robinson, Daniel Rueckert, Sarah Parisot

PII: S1053-8119(17)30302-6
DOI: http://dx.doi.org/10.1016/j.neuroimage.2017.04.014
Reference: YNIMG13955

To appear in: *NeuroImage*
Accepted date: 5 April 2017

Cite this article as: Salim Arslan, Sofia Ira Ktena, Antonios Makropoulos, Emma C. Robinson, Daniel Rueckert and Sarah Parisot, Human Brain Mapping: A Systematic Comparison of Parcellation Methods for the Human Cerebral Cortex *NeuroImage,* http://dx.doi.org/10.1016/j.neuroimage.2017.04.014

Human Brain Mapping: A Systematic Comparison of Parcellation Methods for the Human Cerebral Cortex

Salim Arslan*, Sofia Ira Ktena, Antonios Makropoulos, Emma C. Robinson, Daniel Rueckert, Sarah Parisot

Biomedical Image Analysis Group, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK

Abstract

The macro-connectome elucidates the pathways through which brain regions are structurally connected or functionally coupled to perform a specific cognitive task. It embodies the notion of representing and understanding all connections within the brain as a network, while the subdivision of the brain into interacting functional units is inherent in its architecture. As a result, the definition of network nodes is one of the most critical steps in connectivity network analysis. Although brain atlases obtained from cytoarchitecture or anatomy have long been used for this task, connectivity-driven methods have arisen only recently, aiming to delineate more homogeneous and functionally coherent regions. This study provides a systematic comparison between anatomical, connectivity-driven and random parcellation methods proposed in the thriving field of brain parcellation. Using resting-state functional MRI data from the Human Connectome Project and a plethora of quantitative evaluation techniques investigated in the literature, we evaluate 10 subject-level and 24 groupwise parcellation methods at different resolutions. We assess the accuracy of parcellations from four different aspects: (1) reproducibility across different acquisitions and groups, (2) fidelity to the underlying connectivity data, (3) agreement with fMRI task activation, myelin maps, and cytoarchitectural areas, and (4) network analysis. This extensive evaluation of different parcellations generated at the subject and

*Corresponding author. E-mail address: s.arslan13@imperial.ac.uk

group level highlights the strengths and shortcomings of the various methods and aims to provide a guideline for the choice of parcellation technique and resolution according to the task at hand. The results obtained in this study suggest that there is no optimal method able to address all the challenges faced in this endeavour simultaneously.

Keywords: brain parcellation, resting-state functional MRI, cerebral cortex, functional neuroimaging, model selection, network analysis

1. Introduction

Understanding the brain's behaviour and function has been a prominent and ongoing research subject for over a century (Sporns, 2011). Neuronal interconnections constitute the primary means of information transmission within the brain and are, therefore, strongly related to the way the brain functions (Smith et al., 2013). These connections constitute a complex network that can be estimated at the macro scale via modern imaging techniques such as Magnetic Resonance Imaging (MRI) (Craddock et al., 2013). While structural connectivity networks are typically inferred from diffusion MRI (dMRI), functional networks can be mapped using resting-state functional MRI (rs-fMRI) (Honey et al., 2009; Eickhoff et al., 2015). The former allows estimation of the physical connections, while the latter elucidates putative functional connections between spatially remote brain regions. Analysing brain connectivity from a network theoretical point of view has shown significant potential for identifying organisational principles in the brain and their connections to cognitive procedures and brain disorders (Supekar et al., 2008; Bassett et al., 2008; Smith et al., 2009). This allows to study the brain and its function from a new perspective that accounts for the complexity of its architecture. One of the critical steps in the construction of brain connectivity networks is the definition of the network nodes (Sporns, 2011; Eickhoff et al., 2015). Adopting a vertex- or voxel-based representation yields networks that are very noisy and of extremely high dimensionality, making subsequent network analysis steps often intractable (Thirion

et al., 2014). An alternative approach to node definition is to subdivide the brain into a set of distinct regions - i.e. parcellate the brain-, where each parcel corresponds to a node of the connectivity network.

Traditionally, parcellations derived from anatomical landmarks (e.g. AAL) or cytoarchitectonic information (e.g. Brodmann areas) have been used to define regions of interest (ROIs) for network analysis (Sporns, 2011). Whereas such parcellations are of great importance in order to derive neuro-biologically meaningful brain atlases, they might fail to fully reflect the intrinsic organisation of the brain and capture the functional variability inherent in individual brains, due to brain maturation or injury. Furthermore, they are typically generated on a single or small set of individuals, which can make them biased and unable to accurately represent population variability. This can lead to ill-defined nodes in the constructed network. For example, it has been shown that the anterior cingulate cortex (ACC) exhibits a great amount of heterogeneity in structural (Beckmann et al., 2009) and functional connectivity (Margulies et al., 2007), despite the fact that it is typically represented as a single ROI in a standard anatomical brain atlas (Tzourio-Mazoyer et al., 2002).

Alternatively, random parcellations can be used to define the network nodes. However, this kind of approach could fail to represent the underlying connectivity faithfully and lead to loss of information (Smith et al., 2011). More recent parcellation approaches attempt to overcome these problems by using connectivity information (e.g. rs-fMRI or dMRI data) to drive parcellations (Eickhoff et al., 2015). Since connectivity-based parcellations are directly obtained from the underlying data, such methods can potentially provide highly homogeneous and functionally coherent parcels and separate regions with different patterns of connectivity more accurately. With this idea in mind, several connectivity-driven parcellation methods have been proposed, usually in association with clustering techniques (Thirion et al., 2014). These methods are based on hierarchical clustering (Mumford et al., 2010; Bellec et al., 2010; Arslan and Rueckert, 2015; Moreno-Dominguez et al., 2014), k-means (and its fuzzy counterpart) (Tomassini et al., 2007; Mezer et al., 2009; Golland et al., 2008), Gaus-

sian mixture models (Yeo et al., 2011; Lashkari et al., 2010), spectral graph theory (van den Heuvel et al., 2008; Craddock et al., 2012; Arslan et al., 2015; Parisot et al., 2016a; Shen et al., 2013; Arslan et al., 2016), Markov random fields (MRF) (Ryali et al., 2013; Honnorat et al., 2015; Parisot et al., 2016b), edge detection (Cohen et al., 2008; Laumann et al., 2015; Gordon et al., 2016), region growing (Blumensath et al., 2013; Bellec et al., 2006), independent component analysis (ICA) (Beckmann and Smith, 2004; Smith et al., 2009), Bayesian modelling (Baldassano et al., 2015), meta-analytic connectivity techniques (Eickhoff et al., 2011; Power et al., 2011), dictionary learning (Varoquaux et al., 2011), and many more as extensively reviewed in (Eickhoff et al., 2015; Thirion et al., 2014; de Reus and van den Heuvel, 2013). Although these methods have been thoroughly validated against alternative approaches, a different experimental setup with varying assumptions was used in each case. In addition, the absence of ground truth makes the evaluation of different parcellation methods even more challenging as there is no universally-accepted parcellation that can be used as reference.

In this paper, we propose a systematic comparison of existing parcellation methods using publicly available resources and evaluation measures that are widely used in the literature through a structured experimental pipeline. We focus on resting-state fMRI (rs-fMRI), as the majority of connectivity-driven parcellation methods we are investigating have been developed and tested using this modality. We aim to provide some insight into the reliability of parcellations in terms of reflecting the underlying mechanisms of cognitive function, as well as, revealing their potential impact on network analysis. Thirion et al. (2014) did a similar study at a lower scale, focusing on the analysis of three clustering techniques for fMRI-based brain parcellation, but it only approaches the problem from a clustering point of view. Our study, however, provides a large-scale systematic comparison of the state-of-the-art parcellation methods that encompasses many different aspects in a unified experimental setting.

The main contributions of our study are the following: (1) We evaluate 10 subject-level and 24 groupwise methods using publicly available datasets pro-

vided by the Human Connectome Project (Van Essen et al., 2013b). (2) Our experiments consist of quantitative assessments of parcellations at both subject and group levels and for different resolutions. (3) We evaluate parcellations not only from a data clustering point of view but also with regards to network analysis and multi-modal consistency. Our evaluation includes reproducibility (e.g. Dice coefficient and adjusted Rand index), cluster validity analysis (e.g. Silhouette coefficient and parcel homogeneity) and multi-modal comparisons with task fMRI activation, myelin and cytoarchitectural maps. In addition, we compute network statistics with respect to the underlying parcellation and devise simple network-based tasks (such as gender classification) to evaluate the potential impact of parcellations on network analysis at different scales. It should be noted that our aim is not to directly compare single subject parcellations to group-level ones as groupwise parcellations are subject to methodological biases (e.g. registration) which can affect their performance.

The remainder of this paper is organised as follows: Section 2 summarises the procedures pursued during the generation and evaluation of parcellations. Experimental results are presented in Section 3. In Section 4, we discuss the reliability and applicability of parcellations for network analysis and summarise the impact of this study with some insight into the future of brain parcellation.

2. Materials and Methods

A summary of the processing pipelines is given in Fig. 1. A brief description of subject- and group-level methods is provided in Table 1 and Tables 2-3, respectively. We provide further algorithmic/implementation details for each method in Supplementary Material 1.

2.1. Data

This study is carried out using data from the publicly available Human Connectome Project (HCP) database (Van Essen et al., 2013b), S900 release. All connectivity-driven parcellations are derived from the rs-fMRI acquisitions of

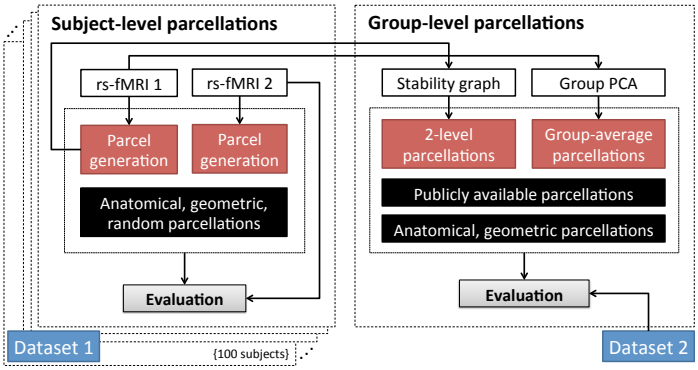


Figure 1: Visual outline of parcellation generation steps for the subject- and group-level parcellations.

100 unrelated subjects (54 female, 46 male adults, aged 22-35). This dataset is publicly available as the "Unrelated 100" in the HCP database and is referred to as "Dataset 1" in the remainder of this paper. For evaluation purposes, we gather a different set of 100 unrelated subjects from the HCP database (Dataset 2) comprising randomly selected 50 male and 50 female adults of age 22-35. The evaluation is performed on Dataset 2 so as to reduce the possible bias towards parcellations computed from Dataset 1 with respect to the provided ones. All subjects had their scans successfully completed for all imaging modalities covered by the HCP.

We use rs-fMRI as our primary data modality for the generation and evaluation of parcellations. This is because most methods selected for this study were developed for rs-fMRI driven parcellation, and rs-fMRI allows test-retest measurements across acquisitions, subjects, and groups. The rs-fMRI scans for each subject were conducted in two sessions, consisting of a total of four runs of approximately 15 minutes each. The sessions were held on different days and during the scans the subjects were presented a fixation cross-hair, projected against a dark background, which prevented them from falling asleep. All subjects were preprocessed by the HCP structural and functional minimal preprocessing pipelines (Glasser et al., 2013). The output of these pipelines for

each subject is a standard set of cortical vertices represented as a triangular mesh $M = \{V, E\}$, where the nodes have a 2 mm spacing. V represents the set of $N = 32k$ nodes, while E describes the connections or edges between neighbouring nodes. This standard mesh is obtained by registering all cortices to a common surface reference space, the Conte69 atlas (Van Essen et al., 2012), using cortical surface based alignment (MSMsulc); implemented using Multimodal Surface Matching (Robinson et al., 2014). Our choice of MSMsulc over the functional based alignment is motivated by the fact that the majority of publicly available parcellations driven by the HCP data used MSMsulc. This yields a set of corresponding mesh coordinates for all subjects. Projection of the 4D rs-fMRI volumes onto the cortical meshes associates each mesh vertex with a rs-fMRI timeseries. Following these preprocessing steps, each timeseries is temporally normalised to zero-mean and unit-variance.

All other modalities are obtained from the HCP dataset (myelin maps, Brodmann areas) (Glasser et al., 2013) or using the HCP processing scripts (task fMRI). Myelin maps are calculated as the ratio of T1-weighted and T2-weighted MRI (Glasser and Van Essen, 2011). The Brodmann parcellation was mapped onto the Conte69 surface atlas (Van Essen et al., 2012) and was then projected onto each subject's cortical surface using the cortical folding driven registration's deformation field. The task fMRI data is preprocessed following the HCP preprocessing pipelines (gradient unwarping, motion and distortion correction, registration to MNI space and projection to the cortical surface). Task activation maps are then obtained using standard FSL tools (FEAT) that use general linear modelling to construct activation maps (Barch et al., 2013). The analysis is carried out separately for each of the 86 different functional contrasts, over 7 different tasks, including the *motor* protocol, the *relational* protocol, the *social* protocol, the *language* protocol, the *emotion* protocol, the *gambling* protocol, and the *working memory* protocol (Barch et al., 2013). We compute the group average myelin maps by averaging all subjects' myelin maps, while the average Brodmann map is obtained with majority voting.

2.2. Parcellation Methods

In order to provide a comprehensive evaluation of the state of the art on surface-based brain parcellation, we gathered 10 single subject and 24 groupwise parcellation methods from the literature. The methods included in this study satisfy at least one of the following criteria:

1. An implementation is publicly available.
2. Pre-computed parcellations are publicly available. Both surface-based and volumetric parcellations are considered.
3. The method can easily be re-implemented.

Subject-level methods

Subject-level methods subdivide the cortical surface of each subject independently. We consider connectivity-driven parcellations as well as anatomical and random parcellations. All single subject methods considered and their associated names used in the remainder of this paper are presented in Table 1. A more detailed description of the different methods is provided in Supplementary Material 1.

Parcellations based on widely used clustering algorithms such as k-means, agglomerative hierarchical clustering (Ward, 1963) and spectral clustering with normalised cuts (Craddock et al., 2012) are computed using in-house implementation built on clustering tools from Python's scikit-learn library and Matlab. The method proposed by Blumensath et al. (2013) is re-implemented as described in the original paper. The remaining connectivity-driven methods (Arslan and Rueckert, 2015; Bellec et al., 2006) are computed using publicly available code.

We also evaluate surface-based anatomical atlases that are distributed as part of the HCP datasets (Desikan et al., 2006; Fischl et al., 2004). These parcellations are tailored to each individual subject with respect to anatomical features, such as cortical folding.

Last but not least, we include two more approaches to our experiments that do not account for any functional data, with the aim of having a baseline towards

Name	Reference	Resolution	Description
Arslan	Arslan and Rueckert (2015); codes available from www.doc.ic.ac.uk/~sa1013/codes.html	Varying	A two-level approach that combines *k*-means and hierarchical clustering. Ward's clustering with Euclidean distance is applied to an initial finer parcellation of 1000 regions per hemisphere.
Blumensath	Blumensath et al. (2013); re-implemented as described in the original paper.	Varying	A two-level method that combines region growing and hierarchical clustering. Ward's clustering with Euclidean distance is applied to an initial finer parcellation of 1000 regions per hemisphere.
Bellec	Bellec et al. (2006); codes available from www.nitrc.org/projects/niak	Varying	A competitive region growing approach driven by parcel homogeneity. A size threshold is applied to avoid over-growing of parcels.
Ward	Ward (1963); in-house implementation, featuring scikit-learn's AgglomerativeClustering function.	Varying	A hierarchical tree is built by merging pairs of clusters, if their similarity is the maximal among the other pairing clusters. Only adjacent clusters are joined into a higher level in order to ensure the spatial contiguity. Clusteringis driven by Ward's linkage rule with Euclidean distance.
K-Means	*k*-means clustering as described in Thirion et al. (2014); in-house implementation, featuring scikit-learn's KMeans and PCA functions.	Varying	PCA is applied to BOLD timeseries for feature reduction. PCA components that explain 50% of the variance combined with spatial vertex coordinates to improve spatial contiguity of parcellations.
N-Cuts	Craddock et al. (2012); in-house implementation of spectral clustering with normalised cuts.	Varying	Spectral clustering with normalised cuts. An affinity matrix is built by correlating the adjacent vertices with each other. Spectral decomposition is applied to the normalised graph Laplacian. The final parcellations are obtained by discretisation.
Destrieux	Fischlet al. (2004); available as part of the HCP datasets from db.humanconnectome.org	150 (75 L, 75 R)	A surface-based parcellation that subdivides the cortex based on the limit between the gyral and sulcal regions.
Desikan	Desikan et al. (2006); available as part of the HCP datasets from db.humanconnectome.org	70 (35 L, 35 R)	Asurface-based parcellation that subdivides the cortex with respect to anatomical landmarks based on the gyri.
Geometric	Geometric parcellations as described in Thirion et al. (2014); in-house implementation, featuring scikit-learn's KMeans function.	Varying	*k*-means clustering is applied to the spatial vertex coordinates. No connectivity information is accounted for.
Random	Random parcellations as described in Schirmer (2015).	Varying	Poisson disk sampling is used to generate regions of approximately equal size by ensuring that tworegion centres are not closer than a given threshold that controls the numberof parcels.

Table 1: Subject-level parcellation methods.

data-driven approaches. We generate (1) random parcellations using Poisson disk sampling as described in (Schirmer, 2015) and (2) geometric parcellations using k-means clustering of the spatial coordinates of the cortex (Thirion et al., 2014).

Two of the single-subject parcellation methods (*Blumensath* and *Arslan*) require an initialisation with a high resolution parcellation (1000 parcels per hemisphere). We use the approach proposed in (Blumensath et al., 2013) to determine the resolution of initial parcellations for each subject, as it is the only method that is driven by seed vertices generated from the underlying data, rather than a set of pre-determined centroids. Using the same initial resolution for both methods facilitates their comparison on a single subject basis.

Group-level methods

Groupwise parcellations build representative models of a population. Methods to obtain a group-level parcellation typically rely on the assumption that

Name	Reference	Resolution	Description
JOINT	Arslan et al. (2015); codes available from www.doc.ic.ac.uk/~sa1013/codes.html	Varying	A surface-based parcellation method based on a joint spectral decomposition of individual subjects. An initial finer parcellation of 2000 regions per hemisphere is used for spatial feature reduction in order to compensate for the computational cost.
2-LEVEL	Groupwise parcellations obtained from the subject-level *Arslan, Blumensath, Bellec, Ward, K-Means*, and *N-Cuts* parcellations.	Varying	2-level approach is applied to the subject-level parcellations of various methods to obtain groupwise parcellations. These parcellations are referred to as the method's name followed by "2" (e.g. Ward-2).
Ward-AVR	Ward (1963); in-house implementation featuring scikit-learn's AgglomerativeClustering function.	Varying	The group average matrix is fed into the Ward's agglomerative hierarchical clustering algorithm using the same setting as for the subject-level *Ward* parcellations.
K-Means-AVR	*k*-means clustering as described in Thirion et al. (2014); in-house implementation featuring scikit-learn's KMeans and PCA functions.	Varying	The group average matrix is fed into *k*-means clustering after being concatenated with the average spatial coordinates to improve spatial contiguity of parcellations.
N-Cuts-AVR	Craddock et al. (2012); in-house implementation of spectral decomposition with normalised cuts.	Varying	A temporal correlation matrix is derived from the group average matrix and transformed into a spatially constrained affinity matrix. Spectral clustering with normalised cuts is used as in the same setting as the subject-level *N-Cuts* parcellations.
GRASP	Honnorat et al. (2015); codes available from cbica.upenn.edu/sbia/software/grasp/index.html	Varying	An MRF-based method that can subdivide the cortex into spatially contiguous parcels by using shape priors. The group average matrix is parcellated into 10000 initial clusters by running the method in the hierarchical clustering mode. Final parcellations are derived from this low-dimensional matrix.
GRAMPA	Parisot et al. (2016b); based on in-house implementation of the method.	Varying	An MRF model that iteratively updates parcel centres and parcel assignments based on modality specific costs. The parcellation is computed using the group average matrix.
Geometric	Geometric parcellations as described in Thirion et al. (2014); in-house implementation featuring scikit-learn's KMeans function.	Varying	*k*-means clustering is applied to the average spatial vertex coordinates. No connectivity information is accounted for.

Table 2: Computed group-level parcellation methods.

spatial correspondence between subjects has been ensured *a priori* by registering subjects to a common template. Hence, each vertex (or voxel) represents the same spatial location for each subject. This allows concatenating or averaging data from different subjects for population-level analysis. The two more popular ways of computing a data-driven groupwise parcellation are (1) performing parcellation for each subject individually and applying a second level clustering algorithm to subject-level parcellations (i.e. *2-level approach*), and (2) computing a representative feature matrix from the population, for instance by concatenating BOLD timeseries across subjects, and submitting this combined matrix to a parcellation method (i.e. *group-average approach*). All the computed and publicly available group-level methods considered in this study are presented in Tables 2 and 3, respectively, along with their associated names. A more detailed description of each method is provided in Supplementary Material 1.

Name	Reference	Resolution	Description
Gordon	Gordon et al. (2016); parcellation available from www.nil.wustl.edu/labs/petersen/Resources.html	333 (161 L, 172 R)	A surface-based parcellation computed from the average gradients of resting-state functional connectivity networks. Provided parcellation is iteratively dilated to cover the entire cortical surface.
Power	Power et al. (2011); parcellation available from balsa.wustl.edu/study/show/WG33	130 (65 L, 65 R)	Resting-state communities originally identified in volume space are projected onto the cortical surface and made publicly available by Van Essen et al. (2017). Connected components within each parcel are relabeled to ensure spatial contiguity.
Yeo	Yeo et al. (2011); parcellation available from balsa.wustl.edu/study/show/WG33	96 (49 L, 47 R)	17-cluster resting-state networks originally derived in volume space from average resting-state functional connectivity data using a GMM-based clustering algorithm are projected onto the cortical surface and made publicly available by Van Essen et al. (2017). Connected components in each parcel are relabeled to ensure spatial contiguity.
ICA	Group-ICA parcellations available from db.humanconnectome.org/data/projects/HCP_500	Varying	Group-average parcellations by means of group-ICA (Beckmann and Smith, 2004) are obtained at several different dimensionalities (25, 50, 100, 200, 300), using a group-PCA output (Smith et al. 2014) from the HCP S500 subjects. Connected components within each parcel are relabeled.
Baldassano	Baldassano et al. (2015); parcellation available from www.princeton.edu/~chrisb/code.html	171 (84 L, 87 R)	A multi-purpose clustering algorithm based on nonparametric Bayesian modeling is applied to dense connectome derived from the HCP S500 group PCA output (Smith et al. 2014) in order to compute a surface-based parcellation.
Glasser	Glasser et al. (2016); parcellation available from balsa.wustl.edu/study/show/RVVG	360 (180 L, 180 R)	A cortical parcellation generated from multi-modal images of 210 adults from the HCP, using a semi-automated approach. Cortical regions are delineated with respect to function, connectivity, cortical architecture, and topography, as well as, expert knowledge and meta-analysis results from the literature.
Fan	Fan et al. (2016); parcellation available from atlas.brainnetome.org	210 (105 L, 105 R)	A volumetric brain parcellation is obtained using both anatomical landmarks and connectivity-driven information. Anatomical regions defined by Desikan et al. (2006) are parcellated into subregions using functional and structural connectivity data from 40 adults provided by the HCP. Cortical parcels are projected from volume to surface.
Shen	Shen et al. (2013); parcellation available from www.nitrc.org/frs/?group_id=51	200 (102 L, 98 R)	A spectral clustering approach is used to compute a volumetric groupwise parcellation based on an optimization process that guarantees functional homogeneity within each parcel and ensures that computed parcels are consistent across subjects. Volumetric parcels from the provided 1 mm sampled 268-parcel atlas are projected to cortical surface.
AAL	Tzourio-Mazoyer et al. (2002); available from www.gin.cnrs.fr/AAL2_files/aal2_for_SPM12.tar.gz	82 (41 L, 41 R)	A popular volumetric atlas that is manually delineated with respect to anatomical landmarks, in particular, by following the sulci course in the brain. Cortical parcels are projected from volume to surface.
Destrieux	Fischl et al. (2004); parcellations available from db.humanconnectome.org	150 (75 L, 75 R)	Majority voting used across subject-level *Destrieux* parcellations to obtain a group-level parcellation.
Desikan	Desikan et al. (2006); parcellations available from db.humanconnectome.org	70 (35 L, 35 R)	Majority voting used across subject-level *Desikan* parcellations to obtain a group-level parcellation.

Table 3: Pre-computed, publicly available group-level parcellation methods.

2-level approach. This technique is similar to majority voting, in the sense that vertices assigned to the same region across subject-level parcellations are clustered together. As a result, group-level parcellations obtained via this method can capture the shared characteristics of the population as approximated by the individual parcellations. To this end, a graphical model of the "parcel stability" is computed across all individual parcellations (Craddock et al., 2012; van den Heuvel et al., 2008). This is achieved by constructing an $N \times N$ adjacency matrix A (i.e. stability graph), in which an edge between vertices v_i and v_j is weighted by the number of times both vertices are assigned to the same parcel

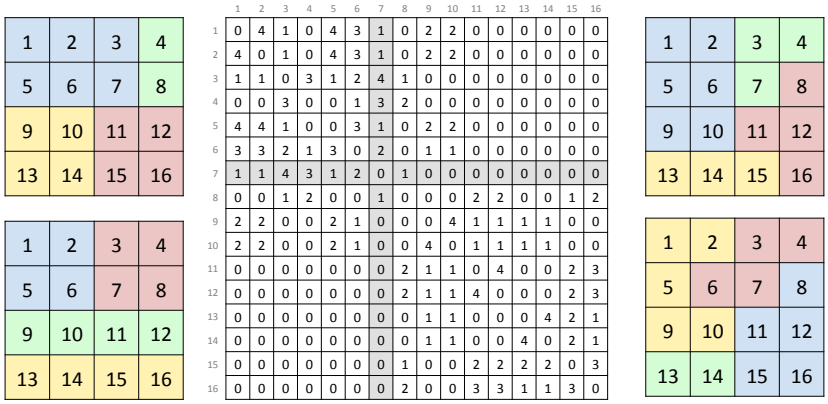


Figure 2: Illustration of how to compute a symmetric 16×16 adjacency matrix A from four toy parcellations (4×4 matrices) where each colour represents a different label/parcel. For example, vertex v_7 (with its corresponding row and column highlighted in A) is assigned to the same parcel as v_3, v_4, and v_6, for 4, 3, and 2 times, respectively, giving $A_{7,3} = 4$, $A_{7,4} = 3$ and $A_{7,6} = 2$. $A_{7,1} = A_{7,2} = A_{7,5} = A_{7,8} = 1$, since it shares the same label with v_1, v_2, v_5, and v_8 in just one parcellation, while the rest of the entries in row 7 of the adjacency matrix are 0, since there does not exist a shared label between the other vertices and v_7 in any of the toy parcellations.

across all individual subject parcellations. Notably, as long as the initial parcellations are spatially contiguous, the spatial integrity of the parcellations is also guaranteed, since only vertices sharing the same cluster membership can be connected in the adjacency matrix. Finally, the graph is subdivided into different number of regions, typically by a graph partitioning algorithm, such as spectral clustering with normalised cuts (van den Heuvel et al., 2008; Craddock et al., 2012). An illustration is provided in Fig. 2 that explains the construction of a stability graph with 4 toy parcellations. This approach is used to generate a group-level parcellation from the individual subject parcellation methods *K-Means*, *Ward*, *N-Cuts*, *Arslan*, *Blumensath*, and *Bellec*.

Group-average approach. This technique aims to capture shared patterns between individuals within a population by computing a group average repre-

sentation of connectivity. This is achieved by concatenating the timeseries of each subject and applying PCA for dimensionality reduction before parcellation (Thirion et al., 2014; Smith et al., 2014). However, using the full-concatenated timeseries with traditional PCA quickly becomes computationally prohibitive when the population's size increases. To overcome this, we follow the methodology employed by the HCP for the generation of group average matrices. We use FSL's incremental group PCA (Smith et al., 2014), a technique developed for computing "pseudo-timeseries" that can (to good approximation) estimate the real PCA output applied to the original combined dataset, while relying on a limited amount of memory.

We apply this technique to generate group level pseudo-timeseries from both Dataset 1 and Dataset 2. Group-level parcellations are computed from each of these datasets using our in-house implementations of clustering techniques (*K-Means*, *Ward* and *N-Cuts*) as well as connectivity-driven methods for which implementations are available (Honnorat et al., 2015; Parisot et al., 2016b).

Other computed parcellations. Alternative to 2-level and group-average approaches, we provide parcellations obtained from a spectral clustering technique that is simultaneously driven by within- and inter-subject connectivity features (Arslan et al., 2015). In addition, a groupwise geometric parcellation is derived using k-means clustering of the average spatial coordinates of all cortical vertices as described in (Thirion et al., 2014).

Publicly available parcellations. Pre-computed, publicly available group-level parcellations are also included in this study (Gordon et al., 2016; Yeo et al., 2011; Power et al., 2011; Baldassano et al., 2015; Fan et al., 2016; Shen et al., 2013; Smith et al., 2014; Glasser et al., 2016; Van Essen et al., 2017). Details on the method and the resolution of the parcellations are provided in Table 3. In particular, it should be noted that the parcellations provided by Baldassano et al. (2015) and the ICA parcellations (Beckmann and Smith, 2004; Smith et al., 2014) are computed from a much larger HCP cohort (group average of

500 subjects) which can comprise our evaluation dataset. This may introduce a bias in the evaluation of both methods.

The methods proposed by Yeo et al. (2011) and Power et al. (2011) as well as the ICA parcellations (Beckmann and Smith, 2004; Smith et al., 2014) were originally developed for identifying communities or resting-state networks (RSNs) that span across the cortical surface, hence do not naturally provide spatially contiguous parcellations. Since it can affect the quality of the evaluation measures, we overcome this by relabelling connected components within each parcel. We then remove very small parcels and slightly dilate the remaining ones to adjust for vertices lost. k-means (both subject-level, 2-level, and group-average versions) and another connectivity-driven approach, *GRAMPA*, can also provide spatially disjoint parcels. In our experiments, we do not apply any post-processing to the parcellations derived by these methods, as we aim to obtain roughly the same number of regions for all computed parcellations for the sake of consistency. Nonetheless, we perform additional experiments to analyse the impact of relabelling connected components for these methods and discuss how their performance changes compared to the original parcellations.

The multi-modal parcellation of the human cerebral cortex (Glasser et al., 2016) is computed through expert manual annotation of imaging data from several modalities, including function, connectivity and cortical architecture. To date, only group level parcellations have been made publicly available, and therefore, we only incorporate this parcellation to our groupwise experiments.

We also include anatomical atlases to our study, including the Automated Anatomical Labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002) and two other parcellations provided by the HCP (Fischl et al., 2004; Desikan et al., 2006). We obtain a groupwise representation of the surface-based anatomical parcellations (Fischl et al., 2004; Desikan et al., 2006) using majority voting across the subject-level parcellations.

Several parcellations are only available in volume space (Tzourio-Mazoyer et al., 2002; Shen et al., 2013; Fan et al., 2016). We use volume-to-surface and surface-to-surface sampling techniques to project volumetric parcels onto the

HCP average cortical atlas (Conte69) (Van Essen et al., 2012). AAL (Tzourio-Mazoyer et al., 2002) and the volumetric parcellation by Shen et al. (2013) are projected onto the cortical surface generated from the Colin27 brain (Holmes et al., 1998) using FreeSurfer (Fischl, 2012), which is then registered to the Conte69 standard space using Multi-modal Surface Matching (Robinson et al., 2014). Our last volumetric parcellation (Fan et al., 2016) is provided in the HCP volumetric space, and is therefore directly projected onto the HCP's standard surface. Finally, all volumetric parcellations are post-processed and each parcel is slightly dilated to fill holes that may have emerged during projection. Unfortunately, volume-to-surface resampling is not a straightforward process, and hence, it is impossible to retain all volume-based parcels after projection. However, we ensure that the parcellation boundaries and relative positions of parcels to each other remain as faithful to the original atlas as possible.

2.3. Parcellation Evaluation Techniques

Evaluating the quality of parcellation methods is a challenging task since there is no ground-truth parcellation of the cerebral cortex. We gather here some of the most commonly used evaluation techniques from the literature to evaluate parcellations at both subject and group levels with respect to varying resolutions. These techniques can be separated into four categories with regards to the parcellation aspects they assess: (1) reproducibility, (2) clustering validity measures, such as homogeneity and Silhouette analysis, (3) multi-modal comparisons with cytoarchitecture, task fMRI activation, and myelination, (4) network analysis. A summary of the evaluation techniques is given in Table 4.

2.3.1. Reproducibility

Reproducibility is a widely-accepted technique for evaluating the robustness of a parcellation method with respect to the underlying data/subjects. It aims at quantitatively measuring the extent of alignment in parcellation boundaries between different parcellations. Reproducibility can be evaluated between parcellations obtained from a) different subjects, b) the same subject but different

Evaluation technique	Description	Quantitative measurements	Level	Previously used in literature
Reproducibility	Assesses the similarity between two sets of parcellations either obtained from different acquisitions of an individual (scan-to-scan) or different groups (group-to-group).	Dice coefficient; Adjusted Rand index	Subject/Group	Craddock et al. (2012); Blumensath et al. (2013); Shen et al. (2013); Thirion et al. (2014); Honnorat et al. (2015); Arslan et al. (2015); Parisot et al. (2016a)
Cluster validity analysis	Evaluates the quality of parcellations from a clustering point of view by measuring the faithfulness of the parcellation to the underlying data source.	Homogeneity; Homogeneity relative to null models; Silhouette coefficient	Subject/Group	Yeo et al. (2011); Craddock et al. (2012); Arslan and Rueckert (2015); Parisot et al. (2016a); Gordon et al. (2016); Arslan et al. (2016)
Agreement with cytoarchitecture	Assesses the overlap with known cytoarchitectonic areas as delineated by the Brodmann atlas.	Dice coefficient	Subject/Group	Blumensath et al. (2013); Arslan et al. (2016); Parisot et al. (2016a)
Goodness-of-fit to task activation	Evaluates how well the parcellations agree with the task activation maps.	Bayesian information criterion	Subject/Group	Thirion et al. (2014); Parisot et al. (2016a)
Alignment with myelination	Assesses the agreement between the parcellations and highly myelinated cortical areas, identified by a coarse myelin-driven parcellation.	Dice coefficient	Subject/Group	Blumensath et al. (2013); Arslan et al. (2016)
Network-based classification	Evaluates the ability of parcellations to capture population differences with a simple gender classification task on functional connectivity networks.	Classification accuracy	Group	Vergun et al. (2013); Satterthwaite et al. (2015)
Graph theoretic analysis	Investigates different topological properties of connectivity networks with a focus on the underlying parcellation.	Clustering coefficient; characteristic path length; small-world index; average node degree	Group	Salvador et al. (2005); Achard et al. (2006)

Table 4: Techniques used to evaluate parcellations.

rs-fMRI acquisitions, c) different groups, and d) different initialisations (if the method depends on the initialisation). Due to the high inter-subject variability within a population, it is not expected to obtain high reproducibility values between different subjects. Nevertheless, a robust parcellation method should yield very similar parcellations for the same subject with different acquisitions. The same should be expected of group level parcellations, assuming the group size is large enough.

We perform a reproducibility analysis for each subject by comparing their parcellations obtained from two different rs-fMRI acquisitions (i.e. scan-to-scan reproducibility). At the group level, we compare the parcellations obtained from Dataset 1 with the ones derived from Dataset 2 (i.e. group-to-group reproducibility). Unfortunately, we are limited to performing the groupwise reproducibility analysis only for the computed parcellations, as only one parcellation/atlas is publicly available from each external source.

Dice coefficient. Dice coefficient (Dice, 1945) is a very popular measure of overlap between two labelled areas. It has been extensively used for evaluating brain parcellations (Craddock et al., 2012; Honnorat et al., 2015; Blumensath et al., 2013; Yeo et al., 2011; Arslan and Rueckert, 2015; Parisot et al., 2016a). Given two parcels X and Y, the Dice coefficient is calculated as:

$$Dice = \frac{2|X \bigcap Y|}{|X| + |Y|}$$

where $|\cdot|$ indicates the number of vertices in a parcel. In order to obtain a global measure of parcellation reproducibility, we follow the approach proposed in (Blumensath et al., 2013). We first compute Dice coefficients for every pair of parcels and match the parcels with the highest overlap. The Dice coefficients of matching parcels are then averaged to obtain a global reproducibility score. The matching process is performed in an iterative manner, where matching pairs identified in one iteration cannot be matched with other parcels at the next iterations. This process is repeated until all pairs are identified. A Dice coefficient of 1 implies a perfect match (identical parcellations).

Low SNR in functional connectivity data or high variability within a group may yield a subdivision of some regions from one parcellation to the next, even when the same algorithm is performed on different acquisitions/subsets. To account for this effect and reduce its impact on reproducibility, we also use a modified version of Dice coefficient that merges the subdivided regions so as to maximise the overlap with the other parcellation as described in (Blumensath et al., 2013). This is done by iteratively matching each parcel in one parcellation to those in the other, if their overlap ratio is ≥ 0.5 (i.e. one parcel comprises at least half of the other parcel). After this process, parcels that are matched with the same parcel are merged and the average Dice coefficient is computed between the matched pairs as described above.

Adjusted Rand index. Adjusted Rand index (ARI) (Hubert and Arabie, 1985) is also considered for the evaluation of parcellation reproducibility (Thirion et al.,

2014). In contrast to Dice coefficient, it measures the agreement of two parcellations without the necessity of initially matching parcels. As a result, it can more effectively measure the agreement between two parcellations with different numbers of clusters (Milligan and Cooper, 1986). The details of the method are given in Appendix A. An ARI of 1 indicates a perfect correspondence between parcellations, whereas a value of 0 implies that the parcellations do not agree on any of the labels.

2.3.2. Cluster validity measures

The second category of validation measures aims to evaluate the similarity of vertices aggregated in the same parcel. Parcellation can be seen as a clustering problem, and there exist many tools targeted at evaluating the quality of clustering solutions. Here, we focus on highly popular measures of clustering quality for brain parcellation, namely parcel homogeneity and Silhouette coefficients. In addition, we adopt the evaluation technique proposed by Gordon et al. (2016) that compares parcellations to a set of "null models" obtained by randomly relabelling the parcellation without altering the relative parcel locations with respect to each other.

Homogeneity. A good parcellation should have the ability to group cortical vertices with highly similar functional connectivity (Craddock et al., 2012; Gordon et al., 2016). It might be particularly important for subsequent network analysis where network nodes are typically represented by the average signal (e.g. BOLD timeseries) within each parcel (Shen et al., 2013; Gordon et al., 2016). This can be evaluated by computing the functional homogeneity within a parcel, a highly popular parcellation measure (Craddock et al., 2012; Shen et al., 2013; Gordon et al., 2016; Arslan and Rueckert, 2015; Parisot et al., 2016a; Honnorat et al., 2015). The homogeneity of a parcel is measured by calculating the average similarity between every pair of vertices assigned to it. This similarity can be defined as the Pearson's correlation coefficient between the "connectivity fingerprints" of vertices (Power et al., 2011; Craddock et al., 2012). A connectivity fingerprint is computed for each vertex v_k, by correlating v_k to the rest of

the cortical vertices and applying Fisher's r-to-z transformation to the resulting correlation coefficients (Power et al., 2011). A global homogeneity value is obtained by averaging the homogeneity values across all parcels (Craddock et al., 2012).

A shortcoming of homogeneity is its dependency on the parcel size distribution. Smaller parcels tend to have a higher homogeneity than large ones, such that, a parcellation mostly composed of many small parcels and a few large regions will perform better than one with similarly sized parcels. To reduce this bias, we compute a weighted arithmetic mean, where each parcel's contribution to the global homogeneity is proportional to its size.

Comparison to null models. While computing homogeneity by means of a weighted mean reduces the bias towards small parcels, homogeneity values remain dependent on the resolution of the parcellations so that fair comparison between different resolutions is not possible. An alternative is proposed in (Gordon et al., 2016) which consists of comparing a parcellation with the so-called "null models" of the same resolution.

In order to obtain such null models, we perform the following procedure: for each hemisphere, we project the parcellation onto a standard spherical surface provided by the HCP and randomly rotate each point in this sphere around the x, y, and z axes. This process moves each parcel to a new location on the cortical surface without altering their relative positions. We then measure the homogeneity of the rotated parcellation and repeat the same process for 1000 different null models. Parcels that move to the medial wall, where no connectivity information is available, are discarded from computations. The advantage of this approach is that it reduces the observed biases with respect to parcel shape and size, as the parcellations are compared to their rotated versions, which have the same resolutions and similar parcel shapes.

In order to quantitatively evaluate parcellations with respect to their null models, we (1) count the number of rotated parcellations with lower homogeneity scores than the original parcellation and (2) compute the difference between

the homogeneity of the original parcellation and the mean homogeneity score of null models, scaled by their standard deviation (i.e. z-scores relative to null models) (Gordon et al., 2016).

Silhouette coefficient. Another useful and popular technique to quantify parcellation reliability is the Silhouette coefficient (SC) (Rousseeuw, 1987), which can be used as an indicator of how well vertices fit in their assigned parcel. For each vertex, it compares the *within-parcel dissimilarity* defined as the average distance to all other vertices in the same parcel, to the *inter-parcel dissimilarity* obtained from those assigned to other parcels (Yeo et al., 2011; Craddock et al., 2012). SC not only evaluates the compactness of parcels, but also their degree of separation from each other. It is defined as follows:

$$SC_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Given a parcellation $\mathbf{U} = \{U_1, U_2, \ldots U_K\}$, a_i and b_i correspond to within-parcel and inter-parcel dissimilarity of vertex $v_i \in U_k$, respectively, and are defined as follows:

$$a_i = \frac{1}{n_k - 1} \sum_{j \in U_k, i \neq j} d(v_i, v_j)$$

$$b_i = \frac{1}{M} \sum_{j \in \mathbb{N}(U_k)} d(v_i, v_j)$$

Here, n_k denotes the number of vertices in U_k, $\mathbb{N}(U_k)$ denotes the set of parcels that are neighbours with U_k, with M being the number of vertices within these neighbouring parcels and $d(v_i, v_j)$ is the distance measure defined as $1-r$, where r is Pearson's correlation computed between the connectivity fingerprints of v_i and v_j. Instead of computing the inter-parcel dissimilarity with respect to the vertices in all other parcels, we restrict the computations to the neighbouring parcels. This is because (1) it is unlikely for a vertex to be assigned to a remote parcel due to spatial constraints imposed on the parcellations, and (2)

computing inter-parcel dissimilarity with respect to all vertices outside a parcel can easily yield a bias towards obtaining high Silhouette coefficients, as the inter-parcel dissimilarity tends to be extremely high due to the many vertices with low similarity contributing to its computation.

Due to the fact that we use correlation distance as the dissimilarity measure, SC ranges within $[-1, +1]$. A negative SC implies misclassification of a vertex, while a value close to 1 indicates that the vertex is clustered with a high degree of confidence. If most vertices possess high Silhouette values, the parcellation is considered to be of high quality. A global score is obtained for each parcellation by averaging the Silhouette coefficients across all vertices.

2.3.3. Comparisons with other modalities

The previously proposed measurements assess the accuracy of parcellations from a clustering point of view. However, when defining regions of interest for neuro-anatomical purposes, the consistency of these areas with well-defined neuro-biological features also constitutes a critical aspect of evaluation. To this end, we expand our comparisons to those with other modalities. We test the parcellation quality by evaluating their agreement with task activation maps, as well as their overlap with myelination patterns and well-known cortical areas delineated from cytoarchitectonic features.

Here, it is worth noting that *Glasser* is not only driven by connectivity, but also uses information from cortical architecture, task fMRI activation, and myelin content. As a result, it may develop a positive bias towards these modalities and this should be taken into consideration while interpreting the performance of *Glasser* with respect to the inter-modality comparisons.

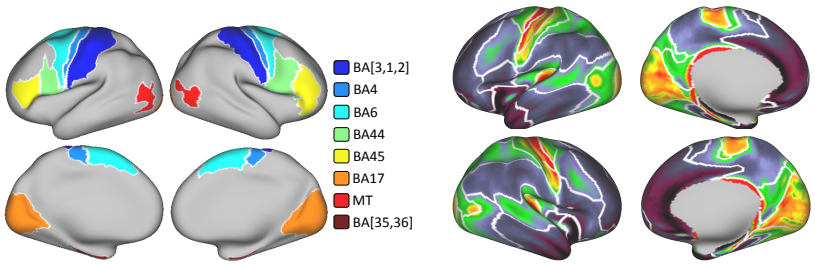
Bayesian information criterion. Bayesian information criterion (BIC) is proposed by Thirion et al. (2014) as a means of quantifying the agreement of parcellations with task fMRI. Each vertex is associated with a task activation map (or the concatenated task activation maps of all subjects if a groupwise parcellation is considered). The BIC criterion measures the goodness-of-fit of a

probabilistic model of the concatenated task activation maps by penalising the negative log-likelihood by the complexity of the model (number of parcels).

Overlap with cytoarchitectonic areas. We measure the agreement of our parcellations with the Brodmann cytoarchitectonic areas (Brodmann, 1909). Although functional connectivity obtained from BOLD timeseries does not necessarily reflect the cytoarchitecture of the cerebral cortex (Wig et al., 2014), agreement with some known cytoarchitectonic areas could indicate a parcellation's ability to reflect the underlying cortical segregation (Gordon et al., 2016). Our standpoint to include comparisons with cytoarchitecture is to show the extent of such agreement at least with certain areas, such as the motor and visual cortex, for which several parcellation techniques report a noticeable alignment (Blumensath et al., 2013; Wig et al., 2014; Gordon et al., 2016). To this end, we use the Brodmann parcellations provided by the HCP, which contain labels for the primary somatosensory cortex (BA 3, 1, and 2), the primary motor cortex (BA 4), the premotor cortex (BA 6), Broca's area (BA 44, 45), the visual cortex (BA 17 and MT), and the perirhinal cortex (BA 35, 36) as shown in Fig. 3.

Quantitative comparisons are performed using the joined Dice coefficient approach as explained in Section 2.3.1. Similarly, overlapping parcels are matched with the Brodmann areas before we compute the Dice coefficient between the matching pairs. It is worth noting that several parcels can be matched to the same area and therefore merged into a larger parcel.

Agreement with structured myelination patterns. Strong similarities have been observed between myelin maps and resting-state fMRI gradients (Glasser and Van Essen, 2011). We should therefore expect the boundaries of rs-fMRI driven parcellations to align with myelination patterns. To evaluate this, we compute a coarse-resolution myelin-driven parcellation (25 parcels) using the method described in (Parisot et al., 2016b) for each subject and a group-level one. This method simply regroups vertices with similar myelin values and, as shown in Fig. 3, effectively delineates the major changes in myelination across the cortex. We compare the parcellations obtained by different methods to these coarse



(a) (b)

Figure 3: Cyto- and myelo-architecture of the cerebral cortex as defined respectively by (a) Brodmann areas and (b) a coarse-resolution myelin-driven parcellation.

parcellations using the joined Dice coefficient approach (Section 2.3.1) which can accurately compare parcellations of different resolutions. Regarding this coarse parcellation we only consider the highly myelinated cortical areas, i.e. cortical areas with a mean myelin value below a certain threshold are discarded.

2.3.4. Network analysis

Parcellations can significantly reduce the dimensionality of the dense human connectome without eliminating valuable information about the interactions between different brain regions and the mechanisms through which these interconnections give rise to complex cognitive processes. It has been common practice in recent neuroscience studies to explore several neurological (Tijms et al., 2013; Fornito et al., 2015) and neuro-developmental disorders (Jafri et al., 2008; Liu et al., 2008; Dennis et al., 2011; Fornito et al., 2012) from a network perspective. These disorders have often been linked to a disruption or abnormal integration of spatially distributed brain regions that would normally be part of a single large-scale network, leading to their characterisation as disconnection syndromes (Catani et al., 2005). Apart from the clinical value of network analysis, efforts to explore potential correlations between connectivity patterns and certain phenotypes like fluid intelligence (Smith, 2016), or to predict an individual's biological age (Robinson et al., 2008; Dosenbach et al., 2010; Pandit et al., 2014; Qiu et al., 2015) have been made. Therefore, a parcellation method can

also be evaluated in terms of its ability to capture the inter-individual variability and to reveal patterns that explain observed cognitive performance.

Once the parcellation has been generated, a network representation can easily be obtained by mapping each network node to a parcel. The edge weights in functional networks usually represent the statistical dependency between the brain regions underlying the connected nodes. In our analysis of functional networks, we use temporal correlation of the representative timeseries as an estimate of the connection strength between two brain parcels.

We explore different ways in which the underlying parcellation can affect network analysis: (1) a network-based classification task and (2) a standard graph theoretic analysis of brain connectivity networks.

Network-based classification. Several studies suggest that differences have been identified in both structural and functional connectivity between genders (Gong et al., 2011). More specifically, in terms of functional connectivity derived from rs-fMRI data, which is the focus of the current survey, significant differences in the topological organisation of functional networks have been found between males and females (Tian et al., 2011). For this reason we choose a gender classification task to evaluate the impact of the parcellation on network-based classification tasks. We use linear Support Vector Machine (SVM) (Burges, 1998), a well-established classifier from the machine learning literature, and a 10-fold cross-validation procedure to get an estimate of each method's performance. Previous studies (Robinson et al., 2008; Vergun et al., 2013; Satterthwaite et al., 2015) have used SVM as a machine learning classifier, which is designed to make predictions based on high-dimensional data, to investigate sex differences in functional connectivity.

Given a set of p-dimensional feature vectors, SVM aims to identify a $(p-1)$-dimensional hyperplane that represents the largest separation or margin between the feature vectors of the two classes. The hyperplane is chosen in a way that the distance from the nearest data point of each class is maximized. The weights assigned to the normalised features to obtain a low-dimensional representation

of the feature vectors can, additionally, be used to describe how heavily weighted the connectivity feature is within the multivariate model (Satterthwaite et al., 2015).

Since node correspondences are ensured with groupwise parcellations, an embedding of each subject's connectivity matrix can be employed to get a general vector representation (Varoquaux and Craddock, 2013), rendering the use of the aforementioned classifier straightforward. This approach is often referred to as "bag of edges" (Craddock et al., 2013) and has been widely used when the underlying parcellation is the same among all subjects.

Graph theoretic analysis. The first step of the analysis involves the computation of partial correlation matrices for all subjects. Partial correlation is considered to discard the "indirect" connections that are preserved by Pearson's correlation, only maintaining the "direct" connections between two regions. It can be computed from the inverse of the empirical covariance matrix, $\mathbf{P} = \Sigma^{-1}$, as $\pi_{vu} = -\mathbf{P}_{vu}(\mathbf{P}_{vv}\mathbf{P}_{uu})^{-1/2}$. It is common practice to perform graph theoretic analysis on partial correlation networks, since correlation coefficient captures the effect of both direct and indirect paths minimising the information added by the analysis (e.g. the shortest path length between two nodes is already captured by full correlation (Fornito et al., 2016)). In order to estimate the group average functional connectivity matrix from the individual partial correlation matrices, we follow the procedure described in (Salvador et al., 2005), where a binary network for the group of subjects can be obtained by testing the null hypothesis that the (mean) partial correlation is zero between any pair i, j of regions. Fisher's r-to-z transform is applied to improve normality and, subsequently, a one-sample t-test is performed for all possible pairs of regions. The False Discovery Rate approach is applied to find the appropriate threshold and to correct for multiple comparisons, according to the steps described in (Benjamini and Yekutieli, 2001), which takes into account the lack of independence between tests. Proportional thresholding is applied after this step, to preserve the top 20% of the edges and reduce threshold effects on the network

measures (Garrison et al., 2015; Alexander-Bloch et al., 2010).

Once the binary network representing the group is generated for each method, a graph theoretic analysis can be performed to investigate topological properties of the network. Network measures of healthy human functional brain connectivity have previously been explored with simple correlation (Eguiluz et al., 2005), partial correlation (Salvador et al., 2005) as well as wavelet correlation (Achard et al., 2006) networks. Here we investigate how some of the most commonly used graph theoretic measures, namely clustering coefficient C, characteristic path length L, their respective normalised versions, γ and λ, as well as the small world index σ, are affected by the underlying parcellation technique and its resolution using partial correlation networks. The details of graph theoretic measures are given in Appendix B.

3. Results

3.1. Experimental Setup

All parcellations included in the subject-level analysis are generated from the individual rs-fMRI scans in Dataset 1. The data for each subject was acquired in two sessions that were held on different days and divided into four runs of approximately 15 minutes each. We concatenate the timeseries of two scans acquired on the same day, obtaining two 30-minute rs-fMRI datasets (rs-fMRI 1 and rs-fMRI 2) for all 100 subjects and use them to obtain two different parcellations for each subject for reproducibility analysis. The groupwise parcellations using the 2-level approach are generated from the individual parcellations obtained from the rs-fMRI 1 set. This set is also submitted to MIGP to obtain the group-PCA matrix, which is subsequently used to compute parcellations using the group-average approach. The rs-fMRI 2 set is exclusively used for the cluster validity measurements (i.e. homogeneity and Silhouette coefficients) of the subject-level parcellations.

Dataset 2 is primarily used to evaluate the groupwise parcellations (publicly available and computed ones from Dataset 1). A second set of group-level

parcellations is also generated using Dataset 2 in order to assess reproducibility across different groups. It is worth noting that, this second set is solely used to assess group-to-group reproducibility and excluded from any other stage of the analysis pipelines.

Most of the pre-computed parcellations comprise a fixed number of regions, while the methods for which an implementation is available can be explored at different resolutions, allowing us to assess the sensitivity of quantitative measures with respect to the number of parcels. For these methods, we generate parcellations containing between 50 and 500 regions (i.e. 25 to 250 per hemisphere), in increments of 50.

Finally, results are reported using the following naming scheme: groupwise parcellations obtained using the 2-level approach will be referred to as their associated method name followed by "2" (e.g. *Ward-2*), whereas parcellations derived from the group-average approach will be accompanied by the "AVR" suffix (e.g. *Ward-AVR*).

3.2. Subject-level Results

For ease of comparison between different methods, we report average evaluation measures in the form of line graphs for all computed resolutions. In order to represent the variability across individuals we show box plots alongside the line graphs, but only for a subset of granularity levels (i.e. for 100, 200, and 300 regions).

Reproducibility results are given in Fig. 4. Cluster validity results, including homogeneity values and Silhouette coefficients, are presented in Fig. 5 and 6, respectively. Bayesian information criterion results obtained from the task activation maps on a per subject basis are given in Fig 8. Finally, comparisons with Brodmann areas and myelin maps are presented in Fig. 9 and 10, respectively.

3.2.1. Reproducibility

Reproducibility results computed by Dice coefficient and adjusted Rand index (ARI) indicate that *Geometric* and *N-Cuts* yield the most reproducible

results. Although *Geometric* shows a better performance than *N-Cuts* at relatively low resolutions, this trend shifts towards the favor of *N-Cuts* at higher resolutions. The performance of *N-Cuts* can be explained by the hard spatial constraints imposed to the adjacency matrices that drive the spectral clustering algorithm, which promotes uniformly sized and/or singleton parcels (Craddock et al., 2012; Blumensath et al., 2013). Obtaining highly reproducible parcellations for *Geometric* is also expected, as the parcellations of a subject are generated from the same set of spatial coordinates.

Two general-purpose clustering methods, *K-Means* and *Ward*, show poor reproducibility scores, in particular when compared to methods derived from an initial finer parcellation such as *Blumensath* and *Arslan*. It is interesting to note that Dice overlap measurements indicate a more favourable performance by *Blumensath* with respect to *Arslan*, while a reverse trend is observed in ARI. These results suggest that methods initialised with a finer parcellation may be more robust, which could be due to the fact that the impact of noise is reduced by the initialisation scheme. *Bellec* generally shows the poorest performance. Nevertheless, it should be noted that this method is originally developed to obtain parcellations with much finer resolutions (over 1000 regions per hemisphere) (Bellec et al., 2006), hence, it may not be adapted to this range of resolutions. Indeed, we can observe that the reproducibility is constantly increasing with the number of parcels. Inversely, the reproducibility of *K-Means* parcellations rapidly decreases with the number of parcels, and *Bellec* surpasses *K-Means* at higher levels of granularity with respect to ARI.

As expected, the Dice coefficient is strongly increased by merging subdivided regions. In particular, this process yields more favourable results for the methods based on hierarchical clustering, namely *Ward*, *Arslan* and *Blumensath*, for which the improvement is up to 15%. *Blumensath* even surpasses *N-Cuts* and *Geometric* over resolutions with more than 150 parcels, becoming the top performing method regarding reproducibility. Other approaches tend to have a less significant improvement, mostly at a rate of $5 - 8\%$, while *N-Cuts* and *Geometric* are minimally affected. This trend can be attributed to the fact that

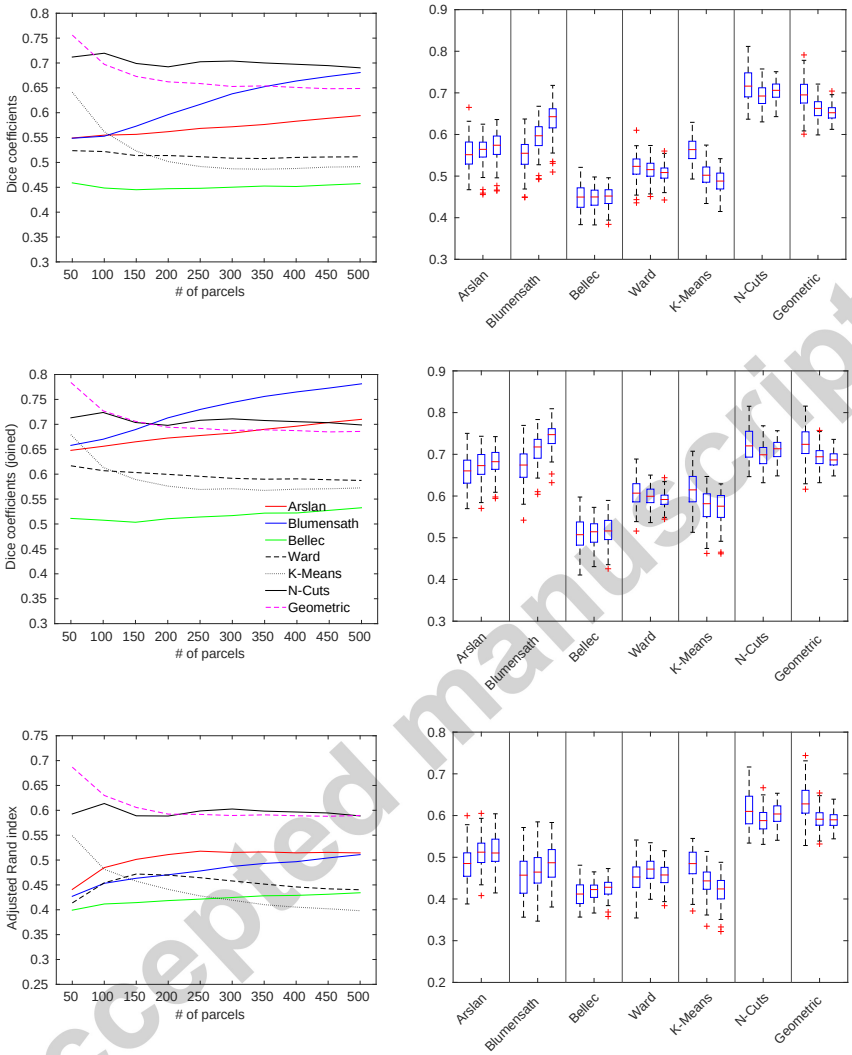


Figure 4: Subject-level reproducibility results. *Left*: Average reproducibility values obtained using Dice coefficient (top), joined Dice coefficient (middle), and adjusted Rand index (bottom). *Right*: Box plots indicate the reproducibility distribution across subjects for 100, 200, and 300 regions, from left to right for each method.

hierarchical clustering subdivides the cortex with a bottom-up process, where boundaries derived at lower resolutions are propagated to higher levels. Joining over-parcellated regions may therefore increase the similarity between parcel-

lations that subdivide the same regions at different levels of the hierarchical clustering tree.

3.2.2. Cluster validity measures

Cluster validity measurements show a clear tendency in favour of connectivity-driven approaches. The most prominent trend is that regardless of the parcellation resolution, *K-Means* outperforms all other methods in terms of both homogeneity (Fig. 5) and Silhouette analysis (Fig. 6). This would indicate that *K-Means* generates the best clustering of the underlying data. It is followed by the hierarchical approaches, each of which performs almost equally regarding Silhouette coefficients, while *Ward* is the best with respect to homogeneity. In particular, *Arslan* consistently generates more homogeneous parcellations than *Blumensath*, which might be attributed to the different techniques used by each method for computing an initial parcellation of the cerebral cortex before applying hierarchical clustering. This initial stage also helps obtain parcellations with a slightly higher degree of confidence than *Ward*. Amongst the connectivity-driven parcellations, *N-Cuts* shows the poorest performance. This can be due to the size bias inherent in this parcellation scheme that could limit the agreement with the underlying data. On the other hand, anatomical parcellations *Desikan* and *Destrieux*, yield the worst measurements and are surpassed by *Geometric* and *Random*. This might suggest that anatomical information alone does not allow to map the brain's functional organisation.

All methods show a performance increasing with the number of parcels computed. This is linked to the fact that both measurements depend on the size of the parcels (e.g. smaller parcels yield better results). It should be noted that this trend may benefit the *K-Means* parcellations, which comprise of several small discontinuous parcels.

Another important observation is the higher inter-subject variability of cluster validity results compared to reproducibility, especially with respect to homogeneity. While one can infer that cluster validity measures are more sensitive than Dice coefficients, this could also be attributed to the fact that reproducibil-

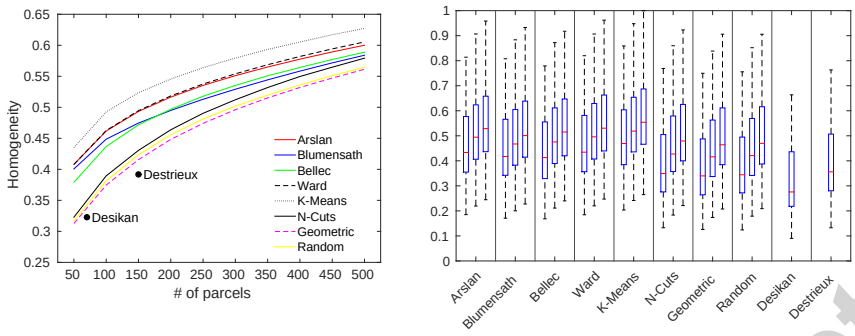


Figure 5: Subject-level homogeneity results. *Left*: Lines show homogeneity values for all resolutions, averaged across subjects, whereas black dots correspond to the average homogeneity obtained from the Desikan and Destrieux atlases, at a fixed resolution of 70 and 150 parcels, respectively. *Right:* Box plots indicate the homogeneity distribution across subjects for 100, 200, and 300 parcels, from left to right for each computed method.

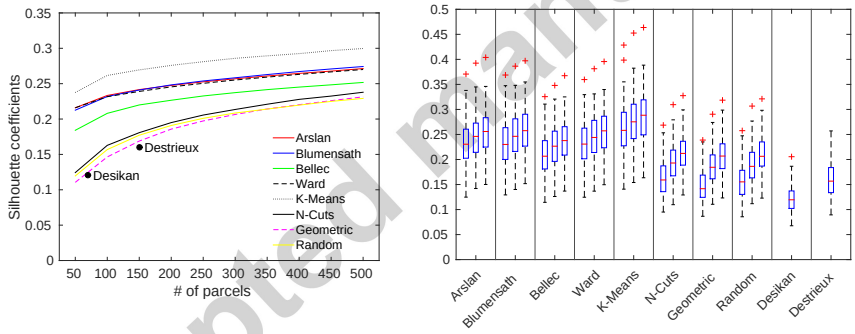


Figure 6: Subject-level Silhouette analysis results. *Left*: Lines show Silhouette coefficients (SC) for all resolutions, averaged across subjects, whereas black dots correspond to the average SC obtained from the Desikan and Destrieux atlases, at a fixed resolution of 70 and 150 parcels, respectively. *Right:* Box plots indicate the SC distribution across subjects for 100, 200, and 300 parcels, from left to right for each computed method.

ity measures the spatial similarity of parcellations that have been registered onto the same standard cortical surface; as a result, an inherent alignment already exists across subjects. This yields a lower inter-subject variability, especially for the spatially constrained methods and with respect to increasing resolution. On the other hand, functional organisation of the brain as estimated by rs-fMRI can

dramatically change from one subject to the next and even between different acquisitions of the same subject. Combining this with the impact of low SNR inherent to rs-fMRI, it may not be possible to parcellate all subjects with high homogeneity and/or confidence. This can be a critical point for consideration, for example, when a group-level study is devised.

Impact of relabelling connected components in disjoint parcellations

K-means clustering applied to the subject-level connectivity data can yield spatially disjoint parcellations. In Fig. 7, we show how certain evaluation measures change when the k-means parcellations are forced to become spatially contiguous by relabelling connected components in each parcel. As can be seen in Fig. 7(a), a large amount of new parcels are generated for all subjects and resolutions after subdividing discontinuous regions. This unsurprisingly yields more homogeneous regions, as homogeneity depends on the resolution and likely to increase when the cortex is parcellated into more subregions (i.e. homogeneous regions still stay homogeneous when subdivided). On the other hand, as we alter the clustering configuration unnaturally by forcing parcels to split, fidelity to the underlying data is negatively affected, yielding lower Silhouette coefficients. Newly generated parcellations provide lower Dice scores at low resolutions, most likely due to the decrease in the overlap ratio between large parcels after splitting. However, it appears that newly obtained (smaller) parcels can be matched better with each other across parcellations, as the joined Dice coefficients and adjusted Rand indices show a more favourable performance after the splitting process.

3.2.3. Multi-modal comparisons

The agreement between the subject-level parcellations and the task fMRI activation maps is evaluated using the Bayesian information criterion (BIC), with respect to all contrasts available in the HCP (a total of 86 activation contrasts from 7 different protocols). The results presented in Fig. 8 show a very similar trend to cluster validity measures, with anatomical parcellations

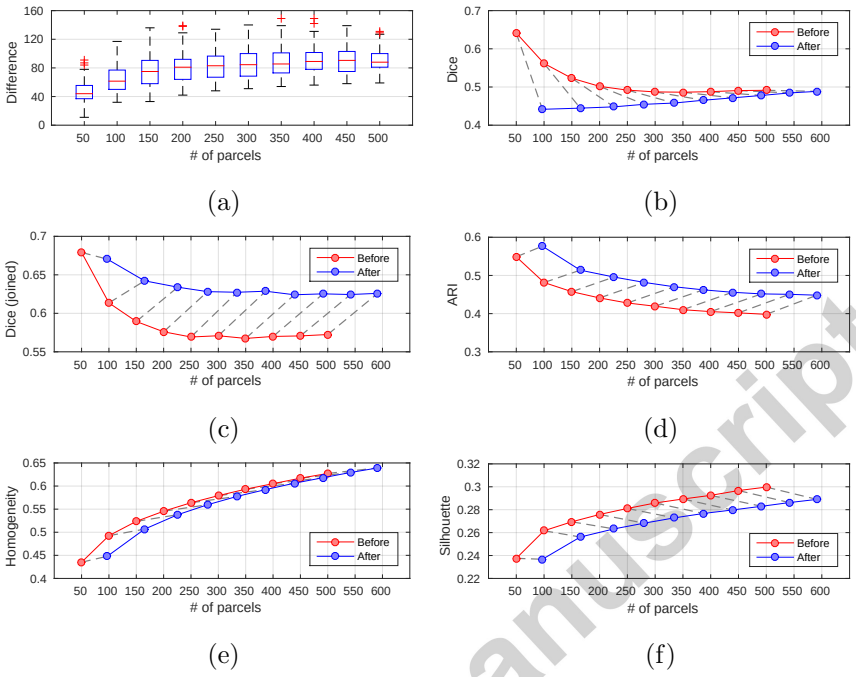


Figure 7: Quantitative evaluation measures obtained from subject-level k-means parcellations, *before* and *after* disjoint parcels are split into spatially contiguous regions. Points representing the original and relabelled parcellations (shown in red and blue, respectively) are matched with dashed lines for ease of comparison. The blue points correspond to the average number of parcels acquired at each resolution after splitting, and therefore, are plotted further to the right with respect to the red points, which align with the resolutions of the original parcellations (50 to 500, in increments of 50) along the x axis. (a) The number of newly generated parcels after splitting, where box plots show the variability across subjects. (b-d) Scan-to-scan reproducibility obtained via Dice similarity, joined Dice similarity, and adjusted Rand index. (e-f) Clustering accuracy measured via parcel homogeneity and Silhouette analysis.

having the worst performance and *K-Means* leading all methods. Interestingly, *Blumensath* has a very poor performance, even being surpassed by *Random* and *Geometric* at high levels of granularity.

The average overlap between the parcellations and the Brodmann areas (BA) across all subjects for all resolutions is given in Fig. 9. In general, all methods have good overlap with the primary somato-sensory cortex (BA[3,1,2]), premotor cortex (BA6), and primary visual cortex (BA17). Relatively low mea-

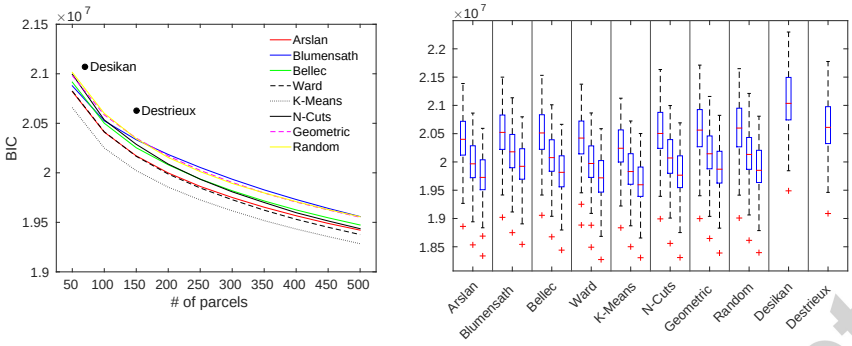


Figure 8: Subject-level Bayesian information criterion (BIC) results showing agreement with task activation. *Left*: Lines show BIC values for all resolutions, averaged across subjects, whereas black dots correspond to the average BIC obtained from the Desikan and Destrieux atlases, at a fixed resolution of 70 and 150 parcels, respectively. *Right:* Box plots indicate the variability across subjects for 100, 200, and 300 parcels, from left to right for each computed method. It should be noted that a lower BIC indicates higher agreement.

sures are obtained for the rest of the Brodmann areas, especially for the perinatal cortex (BA[35,36]). On average, the anatomical parcellations outperform other approaches with the same number of parcels considered, while *N-Cuts* and *Random* yield the best overlap for the rest of the resolutions. On the other hand, *Blumensath* produces the least favourable results at almost all scales.

Average overlap scores obtained by comparing each parcellation with highly myelinated cortical regions are given in Fig. 10. In general, results follow similar trends to those obtained with the comparisons to cyto-architecture. *N-Cuts* and *Random* yield the best overlap scores for all resolutions and anatomical parcellations show a higher degree of agreement with the myelination than the rest of the approaches. Once again, *Blumensath* has the lowest overlap, which might indicate that *Blumensath* parcellations generally do not agree with other cortical features. Similarly, despite its high degree of agreement with task activation, *K-Means* also yields relatively low overlapping scores with both the cyto- and myelo-architecture of the cortex.

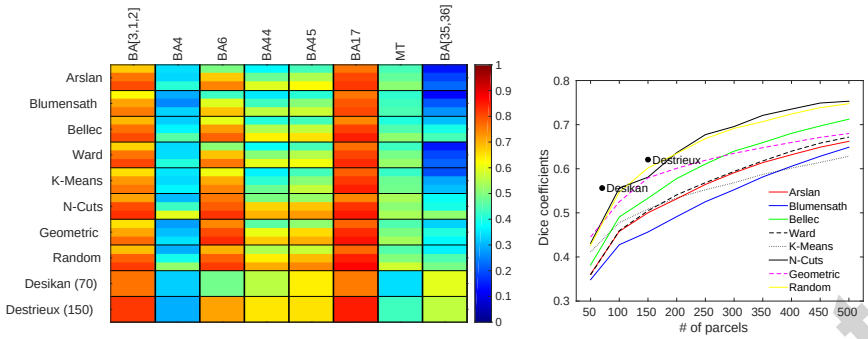


Figure 9: Agreement with the cytoarchitecture of cerebral cortex. *Left*: Overlap of all subject-level parcellations with several Brodmann areas, averaged across individuals. For the connectivity-driven, random and geometric parcellations (top 8 rows), each cell shows overlap scores for 100, 200, and 300 regions, from top to bottom. For the rest of the parcellations, resolutions are indicated aside their names in parentheses. *Right*: Dice coefficients averaged across all considered Brodmann areas for all methods/resolutions

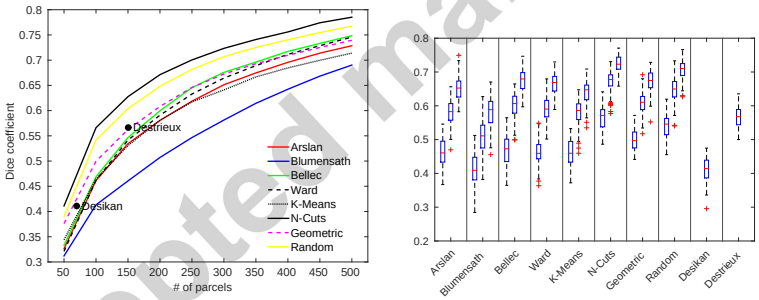


Figure 10: Agreement with the myelo-architecture of the cerebral cortex. *Left*: Dice-based overlap measures of all subject-level parcellations with highly myelinated cortical areas, averaged across individuals. *Right:* Box plots indicate the variability across subjects for 100, 200, and 300 parcels, respectively from left to right for each computed method.

3.3. Group-level Results

Evaluation results obtained for the groupwise parcellations are summarised below.

3.3.1. Reproducibility

The reproducibility values (Fig. 11) are only reported for methods that allow the derivation of multiple parcellations. As expected, the spectral techniques have the best reproducibility results, with *N-Cuts* and *JOINT* leading the others. In general, more favourable results are achieved by 2-level parcellations. This may be attributed to the fact that these parcellations are obtained from a set of individual parcellations that already provide a means of spatial smoothing. Furthermore, the parcellations are computed using normalised cuts, a technique known to increase the reproducibility of parcellations (Craddock et al., 2012; Blumensath et al., 2013). Among the parcellations derived from the average matrix, *Ward-AVR* shows the least favourable performance. MRF-based methods (i.e. *GRASP* and *GRAMPA*) and *K-Means-AVR* also have a relatively poor performance. While joining over-parcellated regions generally increases reproducibility for the group-average approaches, it has a lesser impact on the 2-level parcellations as most of them only show a marginal improvement.

3.3.2. Cluster validity results

Clustering validity results in terms of parcellation homogeneity are summarised in three figures. First of all, homogeneity values obtained by each method/resolution are given in Fig. 12. The homogeneity of each method for a set of selected resolutions together with the homogeneity of their respective null parcellations are presented in Fig. 13. The difference between the homogeneity of the computed parcellations and the distribution of homogeneity models measured as z-scores is shown in Fig. 14. Although group-level homogeneity results are obtained from the average connectivity fingerprints of all subjects, very similar results are achieved when homogeneity is computed on a per subject basis by using each subject's connectivity fingerprints and then averaged across subjects (Supplementary Material 2).

Homogeneity results in Fig. 12 show a relatively poor performance for most of the provided parcellations. The methods that generate the most reproducible parcellations (e.g. spectral methods *JOINT*, *N-Cuts-2*, and *N-Cuts-*

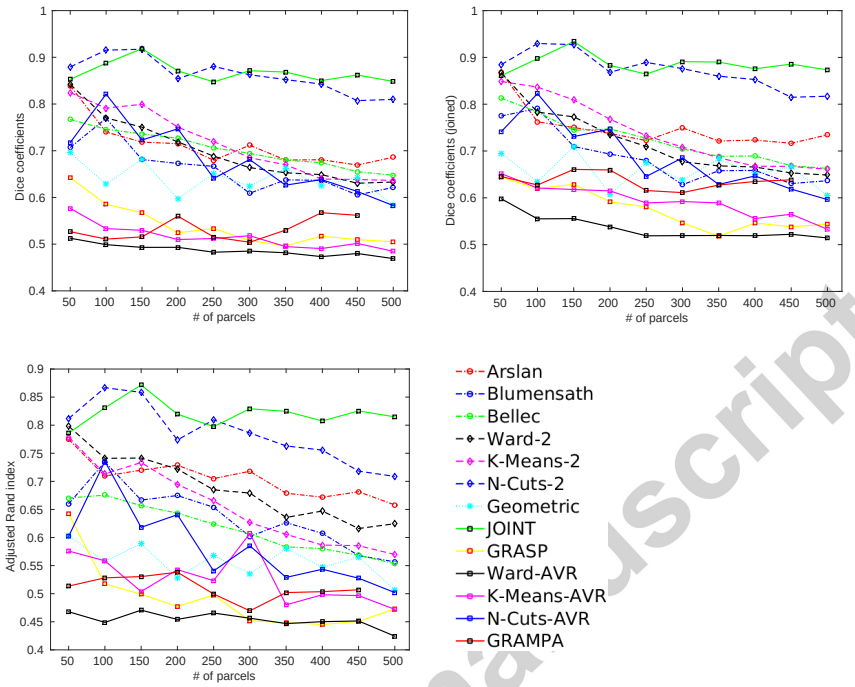


Figure 11: Group-level reproducibility results. Reproducibility values for each method are obtained using Dice coefficient (top, *left*), Dice coefficient after joining over-parcellated regions (top, *right*), and adjusted Rand index (bottom).

AVR) as well as *Geometric* also obtain poor homogeneity values. In general, other connectivity-driven computed parcellations tend to generate highly homogeneous parcellations with the group-average and 2-level methods obtaining very similar results. Among them, *K-Means-AVR* especially excels at lower resolutions, but is outperformed by *Baldassano*, one of the publicly available parcellations based on functional connectivity when similar resolutions are considered. It should be noted, though, that *Baldassano* is obtained from a larger HCP cohort (500 subjects) which may contain our evaluation set and positively bias homogeneity results.

As shown in Fig. 13 and 14, we observe similar performance trends for most of the computed parcellations by comparing to null models. Anatomical parcellations (*AAL*, *Destrieux*, and *Desikan*), and some of the provided parcellations

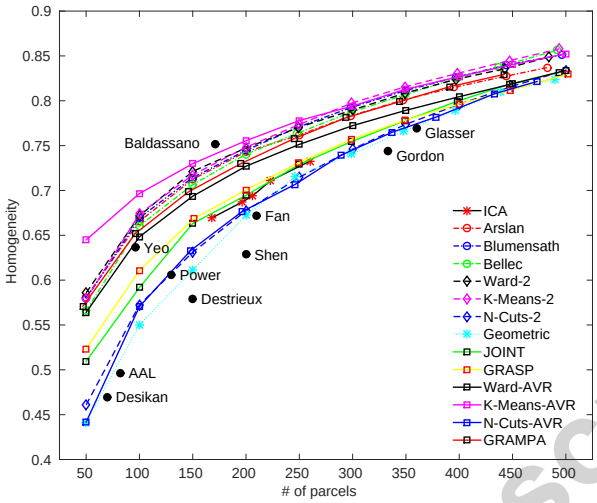


Figure 12: Group-level homogeneity results. Whereas lines show the homogeneity values for all computed resolutions, black dots correspond to the homogeneity scores obtained from the publicly available parcellations with fixed resolutions.

(*Fan*, *Gordon*, and *Shen*), regardless of their respective resolutions perform similar to or worse than their null models. Among the publicly available parcellations, *Baldassano* is on par with *K-Means-AVR*, while *Yeo*, *Power*, and *ICA* also yield good results.

Group-level Silhouette coefficients (Fig. 15) mostly follow the tendency observed in homogeneity. *K-Means-AVR* outperforms the other approaches at all resolutions. It is followed by another group-average technique, *GRAMPA*, which shows a good performance at low levels of granularity. All 2-level approaches, apart from *N-Cuts-2*, perform equivalently well and produce more distinct parcels than most of the group-average methods. In contrast to the homogeneity results, *Gordon* and *Power* are the top-performing provided parcellations. Interestingly, despite producing homogeneous parcellations, *Baldassano*, *Yeo*, and *ICA* show an average performance in terms of Silhouette coefficients. This shows that generating homogeneous parcellations does not necessarily guarantee a good separation between parcels. Overall, spectral techniques perform

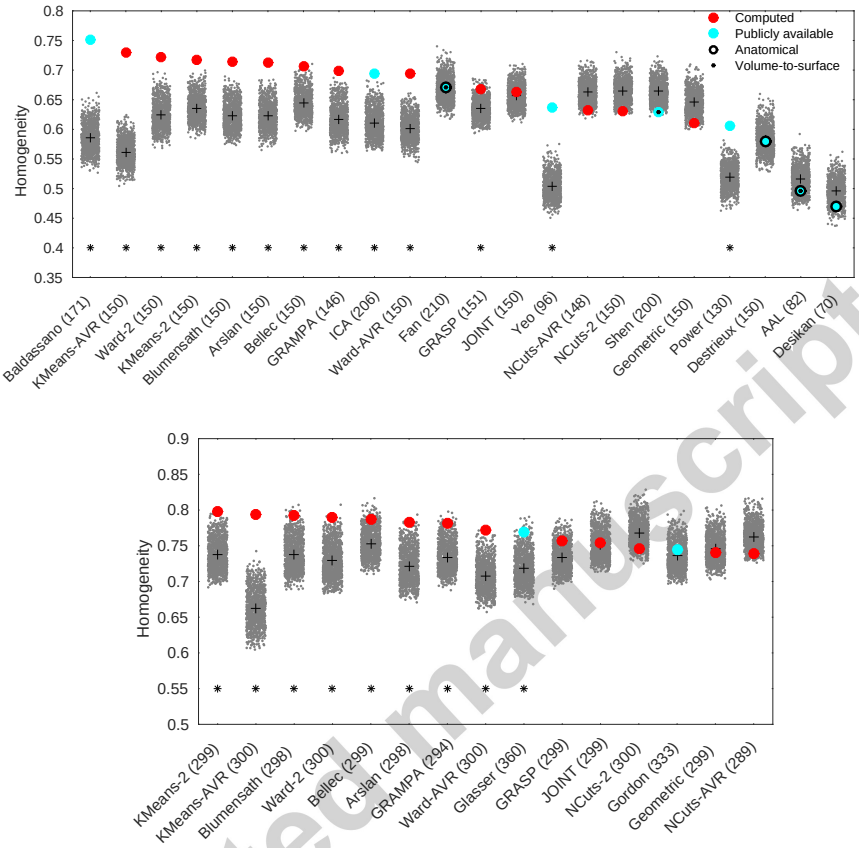


Figure 13: Homogeneity of each parcellation (red and cyan dots) and their respective 1000 null models (gray dots). Null models yield different homogeneity scores due to variation across parcel size and location. + shows the average homogeneity obtained by each set of null parcellations. ∗ indicates that the computed homogeneity is higher than at least 950 of its null parcellations (i.e. $p < 0.05$). *Top:* Results of publicly available parcellations with relatively low resolutions (comprising around or fewer than 200 regions) and the computed parcellations with 150 parcels. *Bottom:* Results of publicly available parcellations with higher resolutions (e.g. comprising around or greater than 300 regions), with the computed parcellations having a fixed resolution of 300 parcels. The exact number of parcels for each method is indicated aside the method name in parentheses.

poorly but still surpass the anatomical and geometric parcellations.

Similarly to homogeneity, we obtain group-level Silhouette coefficients from the average connectivity fingerprints derived from Dataset 2, however, equiva-

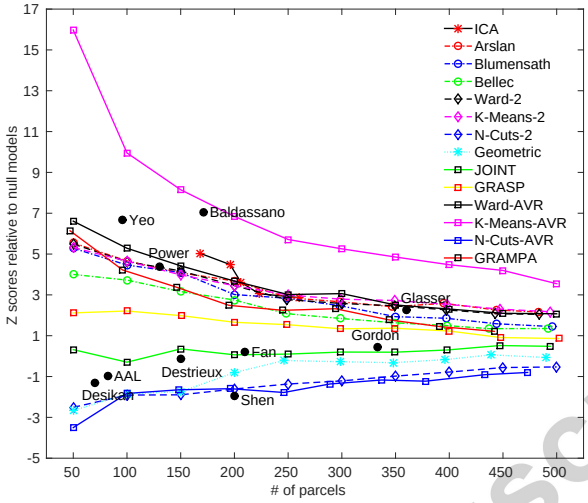


Figure 14: Difference between the actual homogeneity and the homogeneity distribution of null models. Lines show the z scores relative to null models for all computed resolutions, while black dots correspond to the z scores obtained from the publicly available parcellations with fixed resolutions.

lent trends can be observed when Silhouette coefficients are computed for each subject separately and then averaged across subjects (Supplementary Material 2).

Impact of relabelling connected components in disjoint parcellations

Among groupwise parcellation methods, two k-means variants, *K-Means-AVR*, and *K-Means-2* as well as *GRAMPA* can generate spatially disjoint parcellations. In particular, *K-Means-AVR* yields many discontinuous parcels, which significantly increases the total number of parcels after the relabelling process and consequently affects several different evaluation measures as shown in Fig. 16. The changes in reproducibility and cluster validity measures show a similar tendency to those obtained by the subject-level k-means. Although more homogeneous parcels are obtained, the indirect alteration of the clustering configuration leads to lower Silhouette coefficients and z-scores relative to null models. However, this change in the spatial structure of the parcellations in

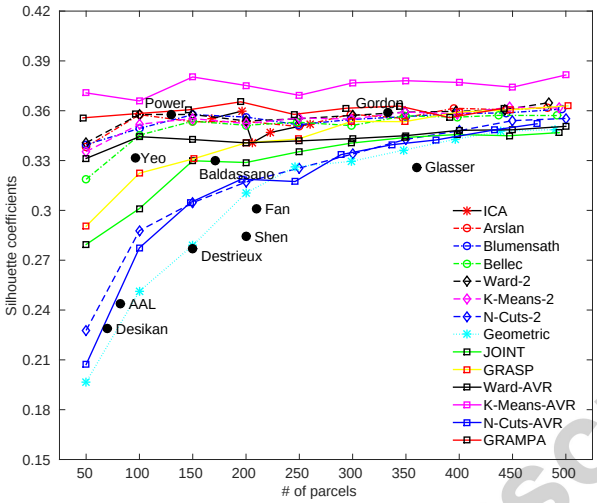


Figure 15: Group-level Silhouette analysis results. Lines show the Silhouette coefficients (SC) for all computed resolutions, while black dots correspond to the SC obtained from the publicly available parcellations with fixed resolutions.

general appears to yield a positive impact on reproducibility, as indicated by the joined Dice coefficients and adjusted Rand indices. The other two methods, *GRAMPA* and *K-Means-2*, only produce few parcels that are discontinuous, thus relabelling does not lead to a significant change in the evaluation measures as shown in Supplementary Material 5.

3.3.3. Multi-modal comparisons

The agreement with concatenated single-subject task activation maps is reported in Fig. 17. In general, all provided parcellations yield relatively poor BIC values compared to the computed parcellations with similar resolutions. The 2-level approaches tend to yield better results than their group-average (AVR) counterparts, in particular at higher resolutions, with *K-Means-2* showing the best performance for most resolutions. This could be linked to the fact that the parcellations are derived from the subject level, where the individual task activation is also estimated from. The only provided methods that show a com-

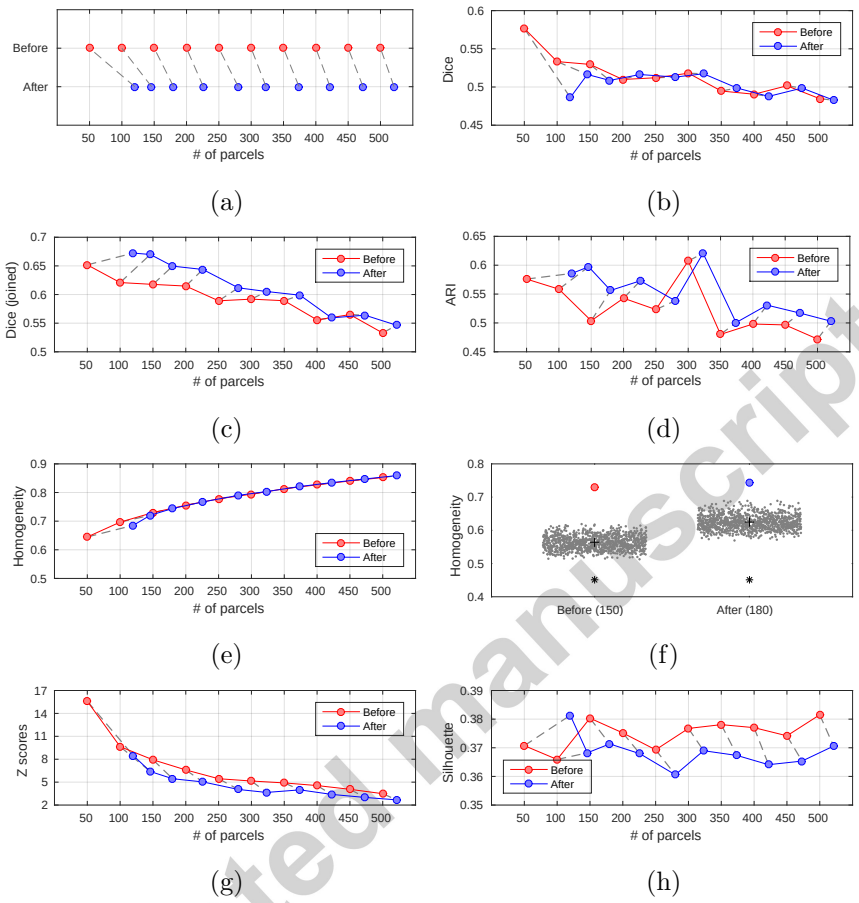


Figure 16: Quantitative evaluation measures obtained from *K-Means-AVR* parcellations, *before* and *after* disjoint parcels are split into spatially contiguous regions. Points representing the original and relabelled parcellations (shown in red and blue, respectively) are matched with dashed lines for ease of comparison. The blue points correspond to the number of parcels acquired at each resolution after splitting, and therefore, are plotted further to the right with respect to the red points, which align with the resolutions of the original parcellations (50 to 500, in increments of 50) along the x axis. (a) The number of parcels before and after the split process. (b-d) Group-to-group reproducibility obtained via Dice similarity, joined Dice similarity, and adjusted Rand index. (e-h) Clustering accuracy measured via parcel homogeneity, comparison to null models (only for one resolution), z scores relative to null models, and Silhouette analysis.

petitive performance are *Yeo* and *Baldassano*, while *GRASP* yields the worst results amongst the computed parcellations. *Glasser* has a poor performance

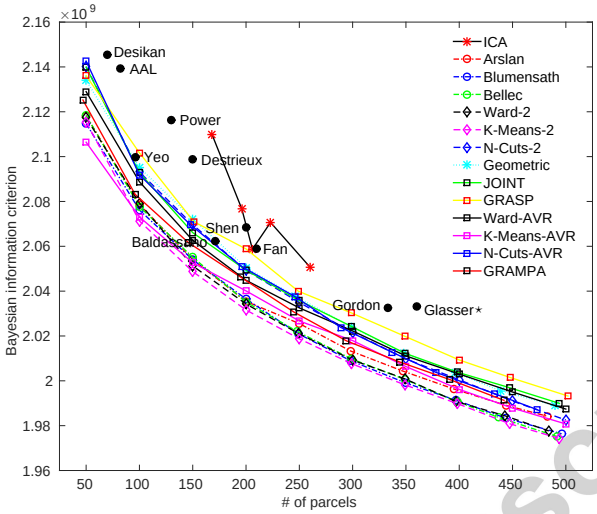


Figure 17: Group-level Bayesian information criterion (BIC) results for measuring the agreement with task activation. Lines show the BIC values for all computed resolutions, while black dots correspond to the BIC obtained from the publicly available parcellations with fixed resolutions. A lower BIC indicates higher agreement with the task activation. ⋆: It should be noted that *Glasser* is partially derived from group average task activation maps, which can influence this evaluation.

despite being driven by task average data. This can be attributed to the fact that it is generated from a different dataset which does not necessarily reflect the single subject task data in our test set.

The overlap between the groupwise parcellations and the average Brodmann areas (BA) for all resolutions is given in Fig. 18. Similarly to the subject-level results, most methods show a high degree of agreement particularly with the primary somato-sensory cortex (BA[3,1,2]), premotor cortex (BA6), and primary visual cortex (BA17). Relatively low measures are obtained for the rest of the Brodmann areas, especially for the perinatal cortex (BA[35,36]). Overall, *Glasser* shows the best performance and yields the highest overlap for most areas. Similarly, other provided parcellations *Fan* and *Gordon*, as well as the anatomical parcellations show a relatively high performance. *Yeo*, *Power*, and *ICA* yield the lowest overlap measures, and in contrast to the general

tendency in the group, do not align well with BA[3,1,2]. Interestingly, *K-Means-AVR* produces the poorest results amongst the computed parcellations for all resolutions. This can be linked to the fact that *K-Means-AVR* parcels are not necessarily spatially contiguous and may be spread across the cortex. In particular, the 2-level approaches perform better than the group-average ones, with *Bellec* leading them at almost all levels of granularity.

Average overlap scores with myelin based parcellations are given in Fig. 19. In general, the 2-level approaches show similar performance and outperform the group-average methods for most resolutions. *Bellec*, *Ward-2* and *K-Means-2* have the highest agreement among the computed parcellations, while *GRAMPA* and *Geometric* yield relatively poor measures. *Glasser* and *Gordon* show the best performance amongst provided parcellations and outperform most of the other approaches when similar resolutions are considered. This is to be expected for *Glasser* since it is partially derived from myelin maps. Other provided parcellations generally yield relatively low measures.

3.3.4. Network analysis

The accuracy achieved for a gender classification task is presented in Fig. 20 and the values of network measures derived with respect to different parcellations/resolutions are shown in Fig. 21.

Connectivity networks are generated using the same set of nodes for all subjects in Dataset 2. The nodes correspond to non-overlapping regions specified by the anatomical atlases, provided parcellations or groupwise data-driven parcellations obtained from Dataset 1. In order to explore the performance of different parcellation methods in capturing population differences, we show the results of a gender prediction task with r-to-z transformed full correlation networks. Before the classification step, dimensionality reduction through PCA is performed (Pereira et al., 2009) and the components explaining 100% of the variance (Robinson et al., 2010) in the training data are preserved for both training and testing. The results with SVM and 10-fold cross-validation are illustrated in Fig. 20. Although there is no single winner across all differ-

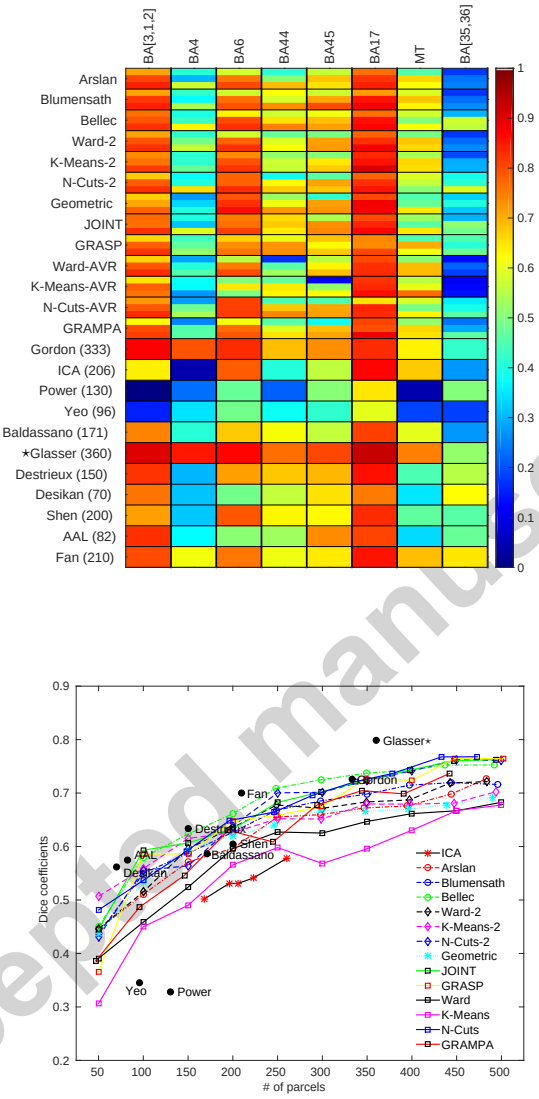


Figure 18: *Top*: Agreement of all group-level parcellations with Brodmann areas. For the computed parcellations (top 13 rows), each cell shows Dice coefficients for 100, 200, and 300 regions, respectively from top to bottom. For the other parcellations, resolutions are indicated aside their names in parentheses. *Bottom*: Average Dice coefficients for each method/resolution. ⋆: It should be noted that *Glasser* uses expert knowledge and priors from the neuro-anatomical literature for the delineation of parcellation borders, which can influence this evaluation.

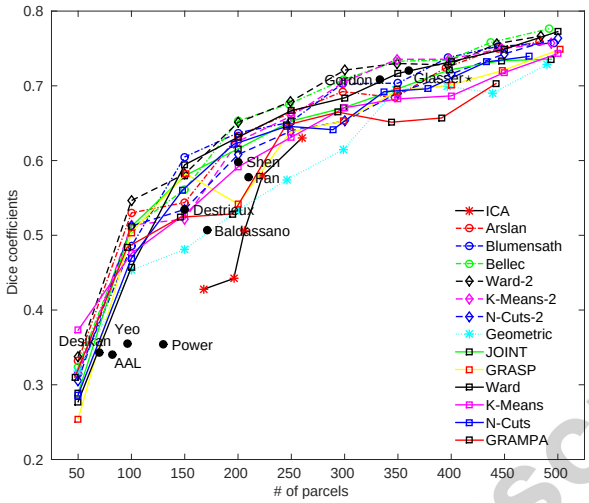


Figure 19: Dice-based overlap measures of all group-level parcellations with highly myelinated cortical areas, derived from a coarse parcellation of the average myelination map. ⋆: It should be noted that *Glasser* is derived from myelin maps and is therefore expected to have a good performance here.

ent resolutions, anatomical parcellations are generally outperformed by several data-driven methods with similar number of parcels. Overall, results obtained with SVM are not very consistent across resolutions, since there is no obvious upward/downward trend with increasing resolution. In fact, most methods demonstrate a similar average performance, being able to classify males and females with above 60% accuracy for granularities below 150 parcels and above 70% for higher resolutions.

More specifically, *Geometric* tends to perform poorly compared to the rest of the methods, both at lower and higher resolutions. The highest SVM classification accuracy (86%) is achieved with *Ward-AVR* and *Glasser* at the scale of 350 and 360 parcels, respectively. Moreover, we can observe that increasing the resolution of the parcellation in data-driven approaches beyond a certain value (350 parcels) does not necessarily provide additional information about population differences. However, lower resolutions lead to lower classification

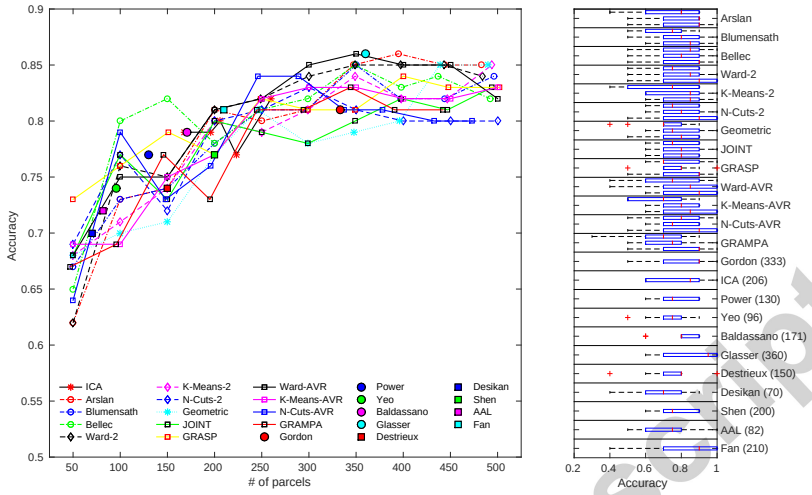


Figure 20: Gender classification results. *Left:* Average Gender classification accuracy on 100 subjects with SVM. *Right:* Variation across results is shown with respect to 10-fold cross-validation.

scores, perhaps due to the fact that functional information valuable for the discrimination between the two classes fades by averaging the signal in larger parcels. Interestingly, *N-Cuts-AVR*, *Bellec* and *Arslan* perform quite well for several resolutions, while *GRASP* yields the top accuracy among all methods for 50 parcels across the cortex. It is also worth mentioning that the parcellations provided by *Yeo*, *Shen* and *Gordon* have below average performance, while *Fan* and *Glasser* have good performance compared to parcellations with similar resolutions.

Our experimental setting allows us to explore the effect of both the parcellation method and the level of granularity on the graph theoretic measures. We compute the most commonly reported network measure in comparative connectomics studies of structural (Sporns et al., 2004; van den Heuvel et al., 2016) and functional (Sporns et al., 2004; Bassett and Bullmore, 2006) brain connectivity, namely, the average clustering coefficient (C), characteristic path length (L), their respective normalised versions, γ and λ (obtained after their division by

the mean corresponding values of a set of 1000 random networks with the same density and degree distribution), the small-world index (σ) and average node degree of the network (k), i.e. the ratio of present connections to the number of nodes divided by 2 (since networks are undirected). The network measures are computed on binarised functional networks obtained by individually testing the elements of the group mean partial correlation matrix (all pairwise connection strengths) for non-zero mean (using $P < 0.01\%$) with correction for multiple comparisons, followed by sparsity-based thresholding (keeping 20% of the edges). Results are presented in Fig. 21.

It can be observed that all six network measures reported are relatively robust with respect to the parcellation method. However, there is an evident effect of parcellation granularity on the calculated measures, which needs to be taken into consideration when performing this kind of network analysis to investigate population differences. This effect is partly attenuated, though, by the use of sparsity-based thresholding that keeps the average node degree k levelled for networks with above than 150 nodes. More specifically, clustering coefficient decreases for resolutions between 100 and 200 nodes to 0.3 and gradually increases to about 0.35 for higher resolutions. On the contrary, γ, its normalised equivalent, increases progressively for resolutions above 150 nodes. The characteristic path length and the normalised λ also increase with resolution, for resolutions above 150 nodes, while a negative trend is observed at lower resolutions. This can be attributed to the fact that k increases with resolution for networks consisting of up to 150 nodes. In general, *GRASP* appears to yield networks with lower γ than the rest of the methods, while *K-Means-AVR* produces networks with higher γ. Similarly to the characteristic path length, the small-world index also increases with resolution with values spanning from 4.7 to 5.3 for the highest resolutions, but always remains above 1 which indicates a small-world topology of functional connectivity networks. The three key measures, γ, λ and small-world index σ demonstrate higher variability between methods at higher resolutions, while they are relatively consistent at low resolutions. Finally, the average node degree increases with resolution up to

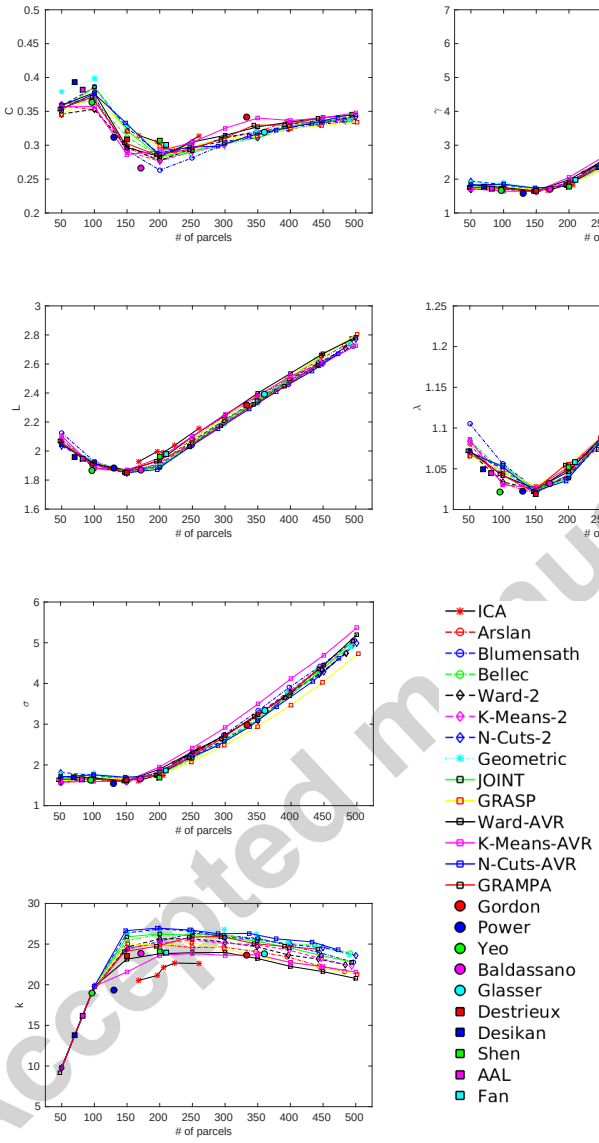


Figure 21: Network measures computed for different parcellations on the group binary networks, including clustering coefficient, C, normalised clustering coefficient, γ, characteristic path length, L, normalised characteristic path length, λ, small-world index, σ and average node degree k.

150 nodes, but remains stagnant afterwards due to fewer connections surviving the significance test after correction for multiple comparisons and proportional thresholding. This has a profound effect on the clustering coefficient and path length.

4. Discussion

In this paper, we presented a large-scale comparison of existing parcellation methods using state-of-the-art evaluation measures and publicly available data provided by the HCP. The generation and evaluation of the parcellations is based on resting-state functional connectivity, which is thought to express the interactions underlying high level cognitive processes. In the absence of a gold standard parcellation, we considered several criteria simultaneously to evaluate the quality of the parcellations, such as reproducibility, parcel homogeneity, and Silhouette analysis. While these measurements assessed the performance from a cluster quality point of view, the neuro-biological interpretation of the obtained parcels is also investigated by comparing parcel boundaries with well-defined neuro-biological properties, such as cytoarchitecture and myelination, as well as task activations. In addition, we devised a simple network analysis task, i.e. gender classification, in order to measure the impact of the underlying parcellation on network analysis, and explored how parcellations affect the structure of connectivity networks based on several network measures.

Our experiments show that there is no clear trend in favour of a specific method - or type of method - regarding all evaluation measures considered. For instance, k-means clustering appears to be largely leading in terms of clustering quality; it, however, shows a poor performance regarding reproducibility and agreement with other modalities. In addition, while cortical delineation intrinsically requires a relatively large number of parcels, this does not appear to be a requirement for effective network analysis. This may suggest that different types of parcellations are to be investigated depending on the task at hand (e.g. one should use different methods when considering network analysis or

cortical delineation). We observe that connectivity-driven parcellations have a much better agreement with the underlying rs-fMRI connectivity compared to anatomical and random parcellations as expected. The benefit of using connectivity to parcellate the brain is not as clear regarding the delineation of cortical areas (agreement with other modalities and established brain delineations) and subsequent network analysis. In particular, anatomical parcellations appear to yield equivalent or better results with respect to cytoarchitecture. A general suggestion regarding network analysis would be to use any parcellation available, since this decision seems to have a very limited impact. However, while this may be true for simple analysis of healthy subjects, it would have to be investigated further in the context of largely different brains (such as diseased subjects or those within a large age range).

Parcellating the cerebral cortex: Aim and scope

The foundations of parcellation were already set in the nineteenth and twentieth centuries, by neuroscientists like Ramon y Cajal, Wernicke, and Brodmann, who emphasised the importance of connectivity in understanding nervous systems and reported insights that underpin the way we think about nervous systems today (Zilles et al., 2010). Although the concept of parcellation spans more than a century in the field of neuroscience and has historically been carried out on the basis of careful studies of the underlying tissue properties, it is currently supplemented with modern *in-vivo* neuroimaging based parcellations (Thirion et al., 2014). The ultimate goal of any kind of parcellation, either based on cytoarchitecture, structural or functional information, is to provide meaningful and homogeneous subdivisions of the brain into regions that are specialised in a certain function. The idea stems from the fact that specific facets of cognition, emotion, and behaviour are considered to be anatomically localised and segregated in the brain. This further allows for a reduction in the complexity of connectivity, an aspect that is highly critical for the study of brain dynamics with whole-brain models.

Therefore, parcellations provide a high-level abstraction of the fundamental

organisation of the brain at macroscopic scales (Sporns et al., 2005; Craddock et al., 2013). Over the last few decades, image acquisition techniques have boosted the potential of *in-vivo* brain mapping and facilitated the multi-scale subdivision of the brain using varying modalities and methods. As a matter of fact, there is not a unique brain parcellation, but rather a spectrum of parcellations that encapsulate fundamental neuro-biological information about cortical organisation and allow the mapping of brain function and anatomy with respect to different aspects. A parcellation may, thus, refer to (1) a reference atlas model that summarises certain properties across the cerebral cortex (e.g. Brodmann atlas, AAL), (2) specialised subunits involved in cognitive functions, (3) high-level structures of functional connectivity (e.g. resting-state networks), or (4) whole-brain subdivisions of the cerebral cortex constituting a few hundred anatomically or functionally distinct parcels (Van Essen et al., 1998; Glasser et al., 2016).

Connectivity estimated from resting-state fMRI and its impact on parcellations

Resting-state fMRI is the most commonly used state-of-the-art technique to map whole-brain functional connectivity, with its high spatial resolution favouring its application over alternative electro-physiological recordings, like EEG and MEG. Its effectiveness to map the function of the brain has been consistently shown across a wide range of studies (Damoiseaux et al., 2006; Salvador et al., 2005; van den Heuvel et al., 2008; Power et al., 2011). However, the true biological interpretation of the BOLD signals is still unknown (Eickhoff et al., 2015), and its low temporal resolution (commonly at the order of seconds) is a limiting factor for the observation of high-frequency patterns. Several sources of noise can influence BOLD signals, including imaging artefacts, head motion, as well as cardiac and respiratory pulsations (Craddock et al., 2013). This, subsequently, leads to a complex connectivity structure, which comprises of linear and nonlinear patterns and is contaminated with noise (Thirion and Faugeras, 2004; Lindquist, 2008). As a consequence, functional connectivity estimated from rs-fMRI usually suffers from false positives and/or indirect connections

mediated by third-party regions (Smith et al., 2011; Eickhoff et al., 2015).

In order to account for the inherently high dimensional and complex structure of the connectivity data, clustering algorithms may *a priori* make various assumptions or introduce implicit/explicit constraints, depending on the task under consideration. This could explain why different parcellation methods perform better or worse with respect to different aspects of the problem. For example, ICA assumes that the fMRI data consists of a mixture of statistically independent components and that spatially distributed functional networks can be effectively separated from signals of non-neural (e.g. artefactual) origin. With a similar objective, but from a different perspective, nonlinear manifold learning techniques rely on the assumption that structures of interest in the connectivity data live in a low dimensional embedding, which can be captured using spectral decomposition (Thirion and Faugeras, 2004; Shen and Meyer, 2006; Langs et al., 2014). Other techniques alter the structure of the connectivity network to obtain more robust parcellations, for instance, by applying thresholding to suppress negative and weak correlations, assuming that correlations under a threshold correspond to spurious connections (van den Heuvel et al., 2008; Power et al., 2011; Craddock et al., 2012; Arslan et al., 2015). It is also common to rely on spatial constraints for computing what is expected to be physiologically more plausible parcellations. Similarly, various methods include a spatial smoothing stage (such as a fine-resolution parcellation) or average subject-level connectivity data for improved SNR and stability in parcellations (Yeo et al., 2011; Blumensath et al., 2013; Arslan and Rueckert, 2015; Gordon et al., 2016). As a general note, it is important to realise that each assumption and processing decision made by a clustering algorithm comes with advantages, as well as limitations, and hence, will inevitably bias the resulting parcellations in different aspects, including the shape, number, size, and spatial contiguity of the parcels (Eickhoff et al., 2015).

Evaluation of parcellations from a clustering point of view

When parcellations are evaluated, both implicit constraints inherent to the method and explicit constraints imposed to the data should be taken into consideration, as they yield inevitable biases towards the computed parcellations (Blumensath et al., 2013). It is, therefore, highly critical to evaluate clustering accuracy from different perspectives.

Hierarchical clustering, k-means, and spectral clustering (as well as their variants) are frequently used to obtain connectivity-driven parcellations, ultimately serving the task of brain mapping (Eickhoff et al., 2015). Their impact on the parcellation configuration as well as their limitations and advantages over each other have been extensively reviewed in (Thirion et al., 2014; Eickhoff et al., 2015). In general, our results align with the previous literature regarding the performance of these clustering algorithms. For example, k-means generally provides the best performing regroupings of the data, but suffers from low reproducibility due to the fact that it does not inherently rely on hard spatial constraints. On the contrary, spectral techniques are usually dominated by spatial constraints, and consequently, capture stable features regarding the geometry of the cortical mesh (Thirion et al., 2014). This appears to confer a strong advantage for reproducibility, but constrains the parcellation task and leads to an inaccurate alignment with the brain's underlying functional organisation. Hierarchical clustering yields a performance that resides in-between: it offers the advantage of generating spatially contiguous parcels, which can contribute to yielding more reproducible parcellations, while still capturing the functional features with high fidelity.

Several other connectivity-driven parcellations computed on a different dataset yield relatively good cluster quality results. One can infer from this observation that similar characteristics shared by healthy adults can be robustly detected across different datasets as long as the analysis is performed on a large cohort (for example *ICA* and *Baldassano* are originally obtained from a group of 500 subjects where this number increases to 1000 for *Yeo*). It should be also noted that, *ICA*, *Baldassano*, and *Glasser* can also comprise some subjects from our

test dataset as they are computed from a larger HCP cohort. This may constitute an important factor promoting a more favourable performance for these methods compared to the others.

Predictably, anatomical parcellations yield the lowest performance in terms of clustering quality. However, they allow a more intuitive neuro-biological interpretation which can make network analysis more insightful. On top of that, our network-based experiments show that a better clustering does not necessarily benefit network analysis. One limitation is their relatively low resolution which is typically addressed by partitioning each parcel into subunits without altering the anatomically delineated boundaries. This can be achieved randomly (Hagmann et al., 2008; Honey et al., 2009) or using functional connectivity (Patel et al., 2008; Fan et al., 2016). The latter approach is adopted by *Fan*, but appears to provide a limited improvement compared to anatomical parcellations.

Agreement of parcellations with other neuro-biological properties of the cortex

The anatomical parcellations based on cortical folding, i.e. *Desikan* and *Destrieux*, as well as the anatomo-functional atlas based on the Desikan parcels (i.e. *Fan*) interestingly show a high degree of agreement with the cytoarchitecture of the cerebral cortex. Although these results may reflect a better alignment between anatomy and cytoarchitectural atlases than with rs-fMRI, this might also be linked to registration errors, as the Brodmann maps are registered to each individual subject based on cortical folding. While we can expect a good overlap in the motor and visual cortex, where the folding patterns are more consistent across subjects, stronger misalignments could occur in other regions.

Similar observations can be made for connectivity-driven parcellations, in which case a higher degree of alignment is found within the motor and visual cortex. Despite the fact that functional connectivity obtained from BOLD time-series is not necessarily expected to reflect the cytoarchitecture of the cerebral cortex, our results agree with several rs-fMRI based studies that report similar findings regarding these regions (Blumensath et al., 2013; Wig et al., 2014; Gordon et al., 2016). On the other hand, a more consistent agreement can be

expected between the connectivity-driven parcellations and highly myelinated areas, as the gradients in rs-fMRI-driven connectivity have been observed to align well with the myelination patterns (Glasser and Van Essen, 2011).

One should also take into consideration the reliability of the evaluation techniques used to compare the different modalities. For example, overlap-based measures, such as the Dice coefficient, are biased by the size of the parcels. Evenly sized/shaped parcels are easier to match with their target parcels, while differences in Dice scores will be much more striking when comparing small parcels over big ones. This bias can lead to more favourable results for some of the parcellations, such as *Geometric*, *N-Cuts*, and *Random*, all of which comprise more uniformly shaped/sized parcels than the rest of the approaches. Although such quantitative measures can provide a means of comparing different methods, the quality of a parcellation with respect to cytoarchitecture or myelin content should also be visually assessed before drawing any conclusion. To this end, we provide visual examples of all the subject-level and groupwise parcellations tested in Supplementary Material 3 and 4, respectively.

Similarly, the Bayesian information criterion has a bias towards more complex models, i.e. parcellations with higher resolution are always favoured (Thirion et al., 2014). It should be also noted that there may exist redundant and contradictory information in the different tasks/contrasts which could bias the results. On top of that, the SNR in the task activation maps is low, therefore, it is likely that the results might be compromised by noise. Finally, our experiments compared group-level parcellations to single subject level task activation maps. While the objective is to evaluate whether these group parcellations provide a good representation of the population, one could also consider comparing to group average task activation maps. This would alleviate single subject noise and could yield better results. For example, the *Glasser* parcellation is expected to have a much better performance with respect to group level task maps on which it is derived.

Additionally, this multi-modal parcellation (*Glasser*) can give a clearer intuition on the behaviour of inter-modality comparisons. This method does not

only rely on resting-state functional connectivity, but also embodies information from task activation, myelin content, and the cortical architecture. It yields very good overlap with the Brodmann areas and myelin content, especially on some parts of the cortex (e.g. motor cortex, highly-myelinated areas), indicating that the overlap measures used for multi-modal comparisons do provide accurate information.

Impact of parcellation on network analysis

Although classification analysis has previously been applied in studies of functional connectivity to predict demographic measures including gender (Satterthwaite et al., 2015; Robinson et al., 2008) and age (Vergun et al., 2013), our experiments suggest that the classification score alone is not a valuable tool for the evaluation of parcellation quality. Instead, the number of features selected (edges in the connectivity matrix) to achieve the same classification performance might be a better means of evaluation provided that a larger number of subjects is available, assuming that a good parcellation should give a sparse selection of features and a more interpretable result. The results obtained with a linear SVM classifier do not favour any particular method, either anatomy, or data driven, to subdivide the brain into regions that would better reflect population differences. In fact, anatomical atlases, like AAL, which are purely based on anatomical landmarks, appear to perform as well as data-driven approaches, designed and tailored to fit the underlying rs-fMRI data. This could be attributed to the specific task at hand, since anatomical and, more specifically, cerebral volume differences have been reported between males and females that significantly influence the volume of white and gray matter (Leonard et al., 2008). Therefore, volume/anatomy- and sex-related differences are hard to disentangle under the current experimental setting, despite the fact that all subjects have been registered to the same anatomical space.

On a different note, the fact that there is no negative effect of higher parcellation resolutions on classification performance indicates that a SVM classifier is appropriate for performing predictions on brain connectivity networks, which

are represented by high-dimensional feature vectors. Interestingly, parcellations producing more evenly sized parcels, like *N-Cuts-AVR*, demonstrate a relative advantage (at least for certain resolutions) over alternative data-driven methods that generate parcels of variable size. According to Stanley et al. (2013), ROIs comprised of more voxels than other ROIs might exhibit greater variability in connectivity, simply due to the fact that a greater variety of signals is included in the ROI itself. As a result, correction mechanisms might be required to account for this variability in parcel extent, which are not required in parcellations consisting of evenly sized parcels. Finally, choosing a different classification task, like disease state or age group, could be more suitable for evaluating parcellation performance in summarising a population's brain connectivity, but the healthy state and narrow age range of the current dataset does not allow this kind of analysis.

As far as graph theoretical analysis is concerned, the measures of network segregation and integration, as well as the small-world topology, seem to be robust to the underlying parcellations. Despite that, all measures are highly susceptible to the granularity of the parcellation (i.e. the number of nodes within the network). These findings align with a previous study on structural connectivity and the sensitivity of network measures to the resolution of the parcellation scheme (Zalesky et al., 2010). The robustness of these network measures to the parcellation method renders them a convenient means for the analysis of population differences and explains their popularity in recent neuroscience studies on healthy and diseased subjects (Wang et al., 2010; Rubinov and Sporns, 2010; Bullmore and Sporns, 2009; Stam et al., 2009, 2007). Nevertheless, the prominent effect of network size on the calculated measures is a factor that needs to be taken into consideration when interpreting the results of relevant studies. To this date, it is difficult to correct for and set limitations to the direct comparison of graph invariants between networks of different order (de Reus and van den Heuvel, 2013). Moreover, the threshold value or significance level chosen to obtain the binarised versions of the functional connectivity networks directly impacts the network density and needs to be reasonably selected and

always accompany the reports of network segregation and integration measures.

Use of parcellations in future subject- and group-level studies

In this paper, evaluations are made on both individual and group level parcellations, with the aim of providing some insight into different techniques that can be used to represent the brain's functional organisation. The results presented as part of this empirical study may indicate parcellation techniques and/or resolutions that are more appropriate for the problem under investigation.

While groupwise parcellations represent shared characteristics within a population, subject-level parcellations serve the purpose of better investigating the functional organisation of an individual brain and understanding the neural basis that results in the observed human behaviour. Evidence suggests that the human connectome possesses connectional traits that are unique to each individual (Mueller et al., 2013; Barch et al., 2013; Wang et al., 2015; Gordon et al., 2017). A recent study (Finn et al., 2015) has further shown that rs-fMRI can be used to derive distinct features to successfully distinguish one individual from another. These features, however, may not be observed in group-averaged datasets (Gordon et al., 2017). Therefore, parcellating the cerebral cortex on a single subject basis can provide a natural starting point for detecting such features, which may further help understand how connectivity varies within a population and how this affects human behaviour and cognition (Wang et al., 2015).

In addition, using subject-level and groupwise parcellations collaboratively may provide more insight into inter-subject variability. For example, cortical regions that are most consistent and/or least similar across subjects can be localised by comparing individual subject parcellations to a group representation obtained via the same clustering method. However, understanding the source of variability across subjects constitutes an additional challenge. While alterations in connectivity can be associated with brain disorders, these could also be attributed to genetic variations (Dubois and Adolphs, 2016), topological differences between subjects (Langs et al., 2014), varying connection strengths

between brain areas in some individuals (Gordon et al., 2017), or even purely caused by registration errors or low SNR in the data. Given many parcellation techniques available at both subject and group levels, analysis of this variability could be an interesting problem to tackle and constitutes one of our planned future directions.

As far as network analysis is concerned, not only the parcellation scheme itself, but its resolution might also have an impact, depending on the task at hand. A recent study suggests that increasing the parcellation resolution yields more reliable biomarkers for studying brain disorders (Abraham et al., 2017). Similarly, using more ROIs for network analysis appears to improve the performance of age prediction tasks (Liem et al., 2017). This might be linked to the fact that parcellations with fewer ROIs may not be able to capture structural patterns of interest from the underlying data due to their resolutions. In this case, data-driven parcellations provide a greater flexibility to study the impact of resolution on network analysis, as they allow the construction of a set of parcellations at different resolutions, as opposed to pre-computed parcellations with fixed resolutions. On a similar note, the heterogeneity of a dataset, e.g. inter-subject variability, could pose additional challenges regarding the performance and interpretability of network analysis. One way to better account for this variability could be generating several group-level parcellations from subsets of the population, preferably on a multi-scale basis, rather than constructing a single parcellation for a population.

Most of the parcellations included in this study can be used to represent the functional organisation of the brain and derive distinct features for network analysis. However, additional information might be required to enhance the information provided by rs-fMRI and identify areas of interest on the cerebral cortex. Evidence suggests that a single modality is too limited to reveal the complex structure of the cerebral cortex, which consists of a mosaic of multiple properties nested at different levels of detail (Glasser et al., 2016; Eickhoff et al., 2015). From a neuro-biological point of view, the integration of other modalities to the parcellation generation task may provide a more accurate and robust

segregation of the cortex, as shown in the recently proposed multi-modal cortical parcellation (Glasser et al., 2016). A prospective future work, therefore, would be to use a similar technique and expand the current evaluation pipeline towards parcellations obtained from different modalities and their combinations.

Limitations

While we did not explore structural connectivity, estimating and analysing brain connectivity from diffusion MRI (dMRI) using tractography techniques is also an important aspect of brain mapping. In contrast to the indirect estimation of connectivity achieved with rs-fMRI, dMRI can estimate the physical white matter connections in the brain. Parcellations derived from dMRI have, therefore, a more intuitive interpretation, and tend to be more robust than rs-fMRI (Parisot et al., 2016a). The estimation of structural connectivity is plagued by several limitations introduced by the imaging technique (a very indirect measurement of white matter connectivity) and processing methods (e.g. tractography) which can suppress existing structural connections, and thus, alleviate the reliability of the connectome analysis. These limitations include the dominance of large fibre bundles, impaired detection of crossing/kissing fibres and long range connections, difficulty to determine the origin or termination of the tracts, and a possible bias with ending tracts in gyri (Van Essen et al., 2013a; Ng et al., 2013). As a result, different tractography algorithms can yield very different estimations of white matter connectivity, while parcellation boundaries tend to align with cortical folding due to this gyral bias. Structural connectivity is, however, a very important aspect of connectomic analysis and parcellations exploiting this modality should be investigated further.

In this empirical study, we considered both surface-based and volumetric parcellations. Whilst efforts are made to be fair to all methods, several important methodological choices have been made, which may have an impact on the evaluation and possibly promote some parcellations over the others. In particular, decisions were made early on to use cortical folding-based alignment to project group parcellations onto individual subjects' functional imaging data. This

choice allows greater consistency with popular volume-based analysis, however it is likely to bias results against groupwise comparisons, especially for comparisons of resting-state homogeneity, and BIC comparisons against task data, where results have consistently shown that resting-state-driven alignment improves the correspondence of resting state, myelin and task across a group (Robinson et al., 2014; Glasser et al., 2016; Sabuncu et al., 2010; Conroy et al., 2013). Furthermore, parcellations are not the products of the same processing pipeline. Most of the publicly available parcellations have been generated under different assumptions, from different sets of subjects with varying cohort size and after being subject to a series of processing steps. Additional processing was further applied to certain methods to make parcellations comparable on a more standard basis. Parcellations that do not naturally provide spatially contiguous cortical regions (e.g. *Yeo*, *Power*, *ICA*) were relabelled, while those that do not cover the entire cortical surface (e.g. *Gordon*) were dilated. Similarly, we used the group-average *Glasser* parcellation in our experiments, despite the fact that this method also provides individual parcellations tailored to each subject. If these subject-specific parcellations are made available, it is likely that their performance with our proposed evaluation measures would see further gains. In particular, the performance of parcellations sampled from a volumetric space should be interpreted carefully due to the complicated transformation steps. Nevertheless, we believe that these parcellations are an essential aspect of our evaluation.

Please see Supplementary Material 3 and 4, for figures showing subject-level and groupwise parcellations used in this study, respectively. All the parcellations included in this paper will be made publicly available via the web page: `http://biomedia.doc.ic.ac.uk/brain-parcellation-survey`, in case one may need access to these parcellations for their own analysis on a different dataset.

Acknowledgements

The research leading to these results has received funding from NIH grant P41EB015902 and the European Research Council under the European Union's Seventh Framework Programme (FP/20072013)/ERC Grant Agreement no. 319456. Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Appendix A: Adjusted Rand Index

ARI is built upon counting the number of items (in our case, vertices) on which two parcellations agree or disagree (Vinh et al., 2009). It classifies $\binom{N}{2}$ pairs of vertices into one of the four sets $(N_{11}, N_{00}, N_{01}, N_{10})$, based on their labeling in each parcellation. For parcellations $\mathbf{U}$ and $\mathbf{V}$, N_{11} corresponds to the number of pairs that are assigned to the same parcel in both $\mathbf{U}$ and $\mathbf{V}$, N_{00} corresponds to the number of pairs that are assigned to different clusters in both $\mathbf{U}$ and $\mathbf{V}$, N_{01} corresponds to the number of pairs that are assigned to the same parcel in $\mathbf{U}$, but different parcels in $\mathbf{V}$, and N_{10} corresponds to the number of pairs that are assigned to the same parcel in $\mathbf{V}$, but different parcels in $\mathbf{U}$. Intuitively, N_{00} and N_{11} account for the agreement of parcellations, whereas N_{01} and N_{10} indicate their disagreement (Vinh et al., 2009). After counting the number of pairs, ARI for parcellations $\mathbf{U}$ and $\mathbf{V}$ is computed as follows:

$$ARI(\mathbf{U}, \mathbf{V}) = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{(N_{00} + N_{01})(N_{01} + N_{11}) + (N_{00} + N_{10})(N_{10} + N_{11})}$$

Appendix B: Graph Theoretical Measures

Graph theory has played an integral role in recent efforts to understand the structure and function of complex systems like the human brain, and has

been widely used to characterise patterns and explore topological properties of connectivity networks. Watts and Strogatz (1998), particularly, focused on two key properties of a network, i.e. the clustering coefficient and the characteristic path length. The clustering coefficient is one of the most elementary measures of local segregation, which measures the density of connections between a node's neighbours. The average of the clustering coefficients for each individual node is the clustering coefficient of the graph. Clustering is significant in a neurobiological context because neuronal units or brain regions that form a densely connected cluster or module communicate a lot of shared information and are therefore likely to constitute a functionally coherent brain system. The clustering coefficient of a binary network can be computed by:

$$C_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k \in N} (a_{ij} a_{jk} a_{ki}) \tag{1}$$

where N is the set of all nodes in the network, k_i is the degree of node i, and a_{ij} is connection status between i and j, with $a_{ij} = 1$ if there is a link and $a_{ij} = 0$ otherwise. The degree of a node is the number of edges attached to it and connecting it to the rest of the network.

While clustering evaluates local connectivity and the segregation of the network into communities, another set of measures captures the capacity of the network to engage in more global interactions that transcend the boundaries of modules and enable network-wide integration. One of the most commonly used measures of integration in brain networks is the characteristic path length, usually computed as the global average of the graph's distance matrix (Watts and Strogatz, 1998). The characteristic path length is a measure of functional integration of the network, demonstrating its ability to quickly combine specialised information from distributed brain regions. A short path length indicates that, on average, each node can be reached from any other node along a path composed of only a few edges. The path length between nodes i and j is given

by:

$$d_{ij} = \sum_{a_{uv} \in g_{i \leftrightarrow j}} a_{uv} \tag{2}$$

where $g_{u \leftrightarrow v}$ is the shortest path between u and v. However, the absolute value of the path length varies greatly with the size and density of individual graphs and, hence, provides only limited information on integration in the network. The network path length should therefore be compared to path lengths of appropriately constructed random networks. For this reason it is customary to compare the obtained path length to that of randomized reference networks with the same number of nodes and edges and identical node degrees as the original network. Such reference networks can be provided by randomizing the original network using a random switching procedure (Rubinov and Sporns, 2010). The calculated values for the clustering coefficient and the path length can, then, be normalised by dividing them with the average corresponding values of the randomized networks. In this study we normalise these metrics using a set of 1000 random networks with the same degree distribution as the original ones.

An important shared feature of complex networks like the human brain is small-world topology (Bullmore and Sporns, 2009). In a small-world network, most links are among neighbouring nodes, but there are a few connections to distant nodes that create shortcuts across the network. As a result, small-world networks are characterised by the prevalence of exquisitely small path lengths among pairs of nodes within very large networks.A prior belief about the small-worldness of the brain arises from the fact that it supports both segregated and distributed information and is also likely evolved to maximise efficiency and minimise the cost of information processing (Bassett and Bullmore, 2006). The small-world index can be calculated as:

$$\sigma = \frac{\gamma}{\lambda} \tag{3}$$

where γ is the normalised clustering coefficient and λ the normalised path length.

References

Abraham, A., Milham, M.P., Martino, A.D., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example. NeuroImage 147, 736 – 745.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., Bullmore, E., 2006. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J Neurosci 26, 63–72.

Alexander-Bloch, A.F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., Lenroot, R., Giedd, J., Bullmore, E.T., 2010. Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. Front Syst Neurosci 4, 147.

Arslan, S., Parisot, S., Rueckert, D., 2015. Joint spectral decomposition for the parcellation of the human cerebral cortex using resting-state fMRI, in: International Conference on Information Processing in Medical Imaging, Springer. pp. 85–97.

Arslan, S., Parisot, S., Rueckert, D., 2016. Boundary mapping through manifold learning for connectivity-based cortical parcellation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 115–122.

Arslan, S., Rueckert, D., 2015. Multi-level parcellation of the cerebral cortex using resting-state fMRI, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 47–54.

Baldassano, C., Beck, D.M., Fei-Fei, L., 2015. Parcellating connectivity in spatial maps. PeerJ 3, e784.

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., et al., 2013. Function in the human connectome: Task-fmri and individual differences in behavior. NeuroImage 80, 169–189.

Bassett, D.S., Bullmore, E., 2006. Small-world brain networks. Neuroscientist 12, 512–523.

Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A., 2008. Hierarchical organization of human cortical networks in health and schizophrenia. J Neurosci 28, 9239–9248.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans Med Imag 23, 137–152.

Beckmann, M., Johansen-Berg, H., Rushworth, M.F., 2009. Connectivity-based parcellation of human cingulate cortex and its relation to functional specialization. J Neurosci 29, 1175–1190.

Bellec, P., Perlbarg, V., Jbabdi, S., Pélégrini-Issac, M., Anton, J.L., Doyon, J., Benali, H., 2006. Identification of large-scale networks in the brain using fMRI. NeuroImage 29, 1231–1243.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multilevel bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51, 1126–1139.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann Stat 29, 1165–1188.

Blumensath, T., Jbabdi, S., Glasser, M.F., Van Essen, D.C., Ugurbil, K., Behrens, T.E., Smith, S.M., 2013. Spatially constrained hierarchical parcellation of the brain with resting-state fMRI. NeuroImage 76, 313–324.

Brodmann, K., 1909. Vergleichende lokalisationslehre der groshirnrinde. Leipzig: Barth 38, 644–645.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Rev Neurosci 10, 186–198.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2, 121–167.

Catani, M., et al., 2005. The rises and falls of disconnection syndromes. Brain 128, 2224–2239.

Cohen, A.L., Fair, D.A., Dosenbach, N.U., Miezin, F.M., Dierker, D., Van Essen, D.C., Schlaggar, B.L., Petersen, S.E., 2008. Defining functional areas in individual human brains using resting functional connectivity MRI. NeuroImage 41, 45–57.

Conroy, B.R., Singer, B.D., Guntupalli, J.S., Ramadge, P.J., Haxby, J.V., 2013. Inter-subject alignment of human cortical anatomy using functional connectivity. NeuroImage 81, 400–411.

Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fMRI atlas generated via spatially constrained spectral clustering. Hum Brain Mapp 33, 1914–1928.

Craddock, R.C., Jbabdi, S., Yan, C.G., Vogelstein, J.T., Castellanos, F.X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., Milham, M.P., 2013. Imaging human connectomes at the macroscale. Nature Methods 10, 524–539.

Damoiseaux, J., Rombouts, S., Barkhof, F., Scheltens, P., Stam, C., Smith, S.M., Beckmann, C., 2006. Consistent resting-state networks across healthy subjects. P Nat Acad Sci 103, 13848–13853.

Dennis, E., Jahanshad, N., Rudie, J., Brown, J., Johnson, K., McMahon, K., de Zubicaray, G., Montgomery, G., Martin, N., Wright, M., et al., 2011.

Altered structural brain connectivity in healthy carriers of the autism risk gene, CNTNAP2. Brain Connectivity 1, 447–459.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.

Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., et al., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn Sci 20, 425–443.

Eguiluz, V.M., Chialvo, D.R., Cecchi, G.A., Baliki, M., Apkarian, A.V., 2005. Scale-free brain functional networks. Phys Rev Lett 94, 018102.

Eickhoff, S.B., Bzdok, D., Laird, A.R., Roski, C., Caspers, S., Zilles, K., Fox, P.T., 2011. Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. NeuroImage 57, 938–949.

Eickhoff, S.B., Thirion, B., Varoquaux, G., Bzdok, D., 2015. Connectivity-based parcellation: Critique and implications. Hum Brain Mapp 36, 4771–4792.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A.R., Fox, P.T., Eickhoff, S.B., Yu, C., Jiang, T., 2016. The human brainnetome atlas: A new brain atlas based on connectional architecture. Cereb Cortex 26, 3508.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprint-

ing: Identifying individuals using patterns of brain connectivity. Nat Neurosci 18, 16641671.

Fischl, B., 2012. Freesurfer. NeuroImage 62, 774–781.

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., et al., 2004. Automatically parcellating the human cerebral cortex. Cereb cortex 14, 11–22.

Fornito, A., Zalesky, A., Breakspear, M., 2015. The connectomics of brain disorders. Nature Rev Neurosci 16, 159–172.

Fornito, A., Zalesky, A., Bullmore, E., 2016. Fundamentals of Brain Network Analysis. Academic Press.

Fornito, A., Zalesky, A., Pantelis, C., Bullmore, E., 2012. Schizophrenia, neuroimaging and connectomics. NeuroImage 62, 2296–2314.

Garrison, K.A., Scheinost, D., Finn, E.S., Shen, X., Constable, R.T., 2015. The (in) stability of functional brain network measures across thresholds. NeuroImage 118, 651–661.

Glasser, M., Coalson, T., Robinson, E., Hacker, C., Harwell, J., Yacoub, E., Ugurbil, K., Anderson, J., Beckmann, C., Jenkinson, M., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. NeuroImage 80, 105–124.

Glasser, M.F., Van Essen, D.C., 2011. Mapping human cortical areas in vivo based on myelin content as revealed by T1- and T2-weighted MRI. J Neurosci 31, 11597–11616.

Golland, Y., Golland, P., Bentin, S., Malach, R., 2008. Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems. Neuropsychologia 46, 540–553.

Gong, G., He, Y., Evans, A.C., 2011. Brain connectivity gender makes a difference. Neuroscientist 17, 575–591.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Dosenbach, N.U., Petersen, S.E., 2017. Individual-specific features of brain systems identified with resting state functional correlations. NeuroImage 146, 918 – 939.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb Cortex 26, 288.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, V.J., Sporns, O., 2008. Mapping the structural core of human cerebral cortex. PLoS Biol 6, e159.

van den Heuvel, M., Mandl, R., Hulshoff Pol, H., 2008. Normalized cut group clustering of resting-state fMRI data. PLoS ONE 3, e2001.

van den Heuvel, M.P., Bullmore, E.T., Sporns, O., 2016. Comparative connectomics. Trends Cogn Sci 20, 345–361.

Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W., Evans, A.C., 1998. Enhancement of MR images using registration for signal averaging. J Comput Assist Tomo 22, 324–333.

Honey, C.J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.P., Meuli, R., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proc Natl Acad Sci USA 106, 2035–2040.

Honnorat, N., Eavani, H., Satterthwaite, T., Gur, R., Gur, R., Davatzikos, C., 2015. GraSP: Geodesic graph-based segmentation with shape priors for the functional parcellation of the cortex. NeuroImage 106, 207–221.

Hubert, L., Arabie, P., 1985. Comparing partitions. J Classif 2, 193–218.

Jafri, M., Pearlson, G., Stevens, M., Calhoun, V., 2008. A method for functional network connectivity among spatially independent resting-state components in schizophrenia. NeuroImage 39, 1666–1681.

Langs, G., Sweet, A., Lashkari, D., Tie, Y., Rigolo, L., Golby, A.J., Golland, P., 2014. Decoupling function and anatomy in atlases of functional connectivity patterns: Language mapping in tumor patients. NeuroImage 103, 462–475.

Lashkari, D., Vul, E., Kanwisher, N., Golland, P., 2010. Discovering structure in the space of fMRI selectivity profiles. NeuroImage 50, 1085–1098.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.Y., Gilmore, A.W., McDermott, K.B., Dosenbach, N.U., Schlaggar, B.L., Mumford, J.A., Poldrack, R.A., Petersen, S.E., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87, 657–670.

Leonard, C.M., Towler, S., Welcome, S., Halderman, L.K., Otto, R., Eckert, M.A., Chiarello, C., 2008. Size matters: Cerebral volume influences sex differences in neuroanatomy. Cereb Cortex 18, 2920–2931.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S.K., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188.

Lindquist, M.A., 2008. The statistical analysis of fMRI data. Stat Sci 23, 439–464.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., Jiang, T., 2008. Disrupted small-world networks in schizophrenia. Brain 131, 945–961.

Margulies, D.S., Kelly, A.C., Uddin, L.Q., Biswal, B.B., Castellanos, F.X., Milham, M.P., 2007. Mapping the functional connectivity of anterior cingulate cortex. NeuroImage 37, 579–588.

Mezer, A., Yovel, Y., Pasternak, O., Gorfine, T., Assaf, Y., 2009. Cluster analysis of resting-state fMRI time series. NeuroImage 45, 1117–1125.

Milligan, G.W., Cooper, M.C., 1986. A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21, 441–458.

Moreno-Dominguez, D., Anwander, A., Knösche, T.R., 2014. A hierarchical method for whole-brain connectivity-based parcellation. Hum Brain Mapp 35, 5000–5025.

Mueller, S., Wang, D., Fox, M.D., Yeo, B.T., Sepulcre, J., Sabuncu, M.R., Shafee, R., Lu, J., Liu, H., 2013. Individual variability in functional connectivity architecture of the human brain. Neuron 77, 586–595.

Mumford, J.A., Horvath, S., Oldham, M.C., Langfelder, P., Geschwind, D.H., Poldrack, R.A., 2010. Detecting network modules in fMRI time series: A weighted network analysis approach. NeuroImage 52, 1465–1476.

Ng, B., Varoquaux, G., Poline, J.B., Thirion, B., 2013. Implications of inconsistencies between fMRI and dMRI on multimodal connectivity estimation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 652–659.

Pandit, A., Robinson, E., Aljabar, P., Ball, G., Gousias, I., Wang, Z., Hajnal, J., Rueckert, D., Counsell, S., Montana, G., et al., 2014. Whole-brain mapping of structural connectivity in infants reveals altered connection strength associated with growth and preterm birth. Cereb Cortex 24, 2324–2333.

Parisot, S., Arslan, S., Passerat-Palmbach, J., Wells, W.M., Rueckert, D., 2016a. Group-wise parcellation of the cortex through multi-scale spectral clustering. NeuroImage 136, 68 – 83.

Parisot, S., Glocker, B., Schirmer, M.D., Rueckert, D., 2016b. GraMPa: Graph-based multi-modal parcellation of the cortex using fusion moves, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 148–156.

Patel, R.S., Borsook, D., Becerra, L., 2008. Modulation of resting state functional connectivity of the brain by naloxone infusion. Brain Imaging Behav 2, 11–20.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: A tutorial overview. NeuroImage 45, S199–S209.

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2011. Functional network organization of the human brain. Neuron 72, 665–678.

Qiu, A., Lee, A., Tan, M., Chung, M.K., 2015. Manifold learning on brain functional networks in aging. Med Image Anal 20, 52–60.

de Reus, M.A., van den Heuvel, M.P., 2013. The parcellation-based connectome: Limitations and extensions. NeuroImage 80, 397–404.

Robinson, E.C., Hammers, A., Ericsson, A., Edwards, A.D., Rueckert, D., 2010. Identifying population differences in whole-brain structural networks: A machine learning approach. NeuroImage 50, 910–919.

Robinson, E.C., Jbabdi, S., Glasser, M.F., Andersson, J., Burgess, G.C., Harms, M.P., Smith, S.M., Van Essen, D.C., Jenkinson, M., 2014. MSM: A new flexible framework for multimodal surface matching. NeuroImage 100, 414–426.

Robinson, E.C., Valstar, M., Hammers, A., Ericsson, A., Edwards, A.D., Rueckert, D., 2008. Multivariate statistical analysis of whole brain structural networks obtained using probabilistic tractography, in: International Conference

on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 486–493.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20, 53–65.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52, 1059–1069.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2013. A parcellation scheme based on von Mises-Fisher distributions and markov random fields for segmenting brain regions using resting-state fMRI. NeuroImage 65, 83–96.

Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., Haxby, J.V., 2010. Function-based intersubject alignment of human cortical anatomy. Cereb Cortex 20, 130–140.

Salvador, R., Suckling, J., Coleman, M.R., Pickard, J.D., Menon, D., Bullmore, E., 2005. Neurophysiological architecture of functional magnetic resonance images of human brain. Cereb Cortex 15, 1332–1342.

Satterthwaite, T.D., Wolf, D.H., Roalf, D.R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E.D., Elliott, M.A., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R.E., Gur, R.C., 2015. Linked sex differences in cognition and functional connectivity in youth. Cereb Cortex 25, 2383.

Schirmer, M.D., 2015. Developing brain connectivity: Effects of parcellation scale on network analysis in neonates. Ph.D. thesis. King's College London.

Shen, X., Meyer, F.G., 2006. Nonlinear dimension reduction and activation detection for fMRI dataset, in: Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, IEEE. pp. 90–90.

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. NeuroImage 82, 403–415.

Smith, S., 2016. Linking cognition to brain connectivity. Nature Neurosci 19, 7–9.

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., et al., 2009. Correspondence of the brain's functional architecture during activation and rest. P Natl Acad Sci 106, 13040–13045.

Smith, S.M., Hyvärinen, A., Varoquaux, G., Miller, K.L., Beckmann, C.F., 2014. Group-PCA for very large fmri datasets. NeuroImage 101, 738–749.

Smith, S.M., Miller, K.L., Salimi-Khorshidi, G., Webster, M., Beckmann, C.F., Nichols, T.E., Ramsey, J.D., Woolrich, M.W., 2011. Network modelling methods for fMRI. NeuroImage 54, 875–891.

Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., Barch, D.M., Uurbil, K., Essen, D.C.V., 2013. Functional connectomics from resting-state fMRI. Trends Cogn Sci 17, 666–682.

Sporns, O., 2011. The human connectome: A complex network. Ann N Y Acad Sci 1224, 109–125.

Sporns, O., Chialvo, D.R., Kaiser, M., Hilgetag, C.C., 2004. Organization, development and function of complex brain networks. Trends Cogn Sci 8, 418–425.

Sporns, O., Tononi, G., Ktter, R., 2005. The human connectome: A structural description of the human brain. PLOS Comp Bio 1, e42.

Stam, C., De Haan, W., Daffertshofer, A., Jones, B., Manshanden, I., Van Walsum, A.V.C., Montez, T., Verbunt, J., De Munck, J., Van Dijk, B., et al., 2009. Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease. Brain 132, 213–224.

Stam, C., Jones, B., Nolte, G., Breakspear, M., Scheltens, P., 2007. Small-world networks and functional connectivity in Alzheimer's disease. Cereb cortex 17, 92–99.

Stanley, M.L., Moussa, M.N., Paolini, B., Lyday, R.G., Burdette, J.H., Laurienti, P.J., 2013. Defining nodes in complex brain networks. Front Comput Neurosci 7, 169.

Supekar, K., Menon, V., Rubin, D., Musen, M., Greicius, M.D., 2008. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. PLoS Comput Biol , 1–11.

Thirion, B., Faugeras, O., 2004. Nonlinear dimension reduction of fMRI data: the laplacian embedding approach, in: Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on, IEEE. pp. 372–375.

Thirion, B., Varoquaux, G., Dohmatob, E., Poline, J.B., 2014. Which fMRI clustering gives good brain parcellations? Front Neurosci 8, 167.

Tian, L., Wang, J., Yan, C., He, Y., 2011. Hemisphere-and gender-related differences in small-world brain networks: a resting-state functional MRI study. NeuroImage 54, 191–202.

Tijms, B., Möller, C., Vrenken, H., Wink, A., de Haan, W., van der Flier, W., Stam, C., Scheltens, P., Barkhof, F., 2013. Single-subject grey matter graphs in Alzheimer's disease. PloS ONE 8, e58921.

Tomassini, V., Jbabdi, S., Klein, J.C., Behrens, T.E., Pozzilli, C., Matthews, P.M., Rushworth, M.F., Johansen-Berg, H., 2007. Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations. J Neurosci 27, 10259–10269.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling

of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289.

Van Essen, D.C., Drury, H.A., Joshi, S., Miller, M.I., 1998. Functional and structural mapping of human cerebral cortex: Solutions are in the surfaces. P Natl Acad Sci 95, 788–795.

Van Essen, D.C., Glasser, M.F., Dierker, D.L., Harwell, J., Coalson, T., 2012. Parcellations and hemispheric asymmetries of human cerebral cortex analyzed on surface-based atlases. Cereb Cortex 22, 2241–2262.

Van Essen, D.C., Jbabdi, S., Sotiropoulos, S.N., Chen, C., Dikranian, K., Coalson, T., Harwell, J., Behrens, T.E., Glasser, M.F., 2013a. Mapping connections in humans and nonhuman primates: aspirations and challenges for diffusion imaging. Diffusion MRI, 2nd edition (eds. Johansen-Berg, H. & Behrens, TEJ) , 337–358.

Van Essen, D.C., Smith, J., Glasser, M.F., Elam, J., Donahue, C.J., Dierker, D.L., Reid, E.K., Coalson, T., Harwell, J., 2017. The brain analysis library of spatial maps and atlases (BALSA) database. NeuroImage 144, 270–274.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., 2013b. The WU-Minn Human Connectome Project: An overview. NeuroImage 80, 62 – 79.

Varoquaux, G., Craddock, R.C., 2013. Learning and comparing functional connectomes across subjects. NeuroImage 80, 405–415.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B., 2011. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity, in: Biennial International Conference on Information Processing in Medical Imaging, Springer. pp. 562–573.

Vergun, S., Deshpande, A., Meier, T., Song, J., Tudorascu, D., Nair, V., Singh, V., Biswal, B., Meyerand, M., Birn, R., Prabhakaran, V., 2013. Characteriz-

ing functional connectivity differences in aging adults using machine learning on resting state fMRI data. Front Comput Neurosci 7, 38.

Vinh, N.X., Epps, J., Bailey, J., 2009. Information theoretic measures for clusterings comparison: Is a correction for chance necessary?, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM. pp. 1073–1080.

Wang, D., Buckner, R.L., Fox, M.D., Holt, D.J., Holmes, A.J., Stoecklein, S., Langs, G., Pan, R., Qian, T., Li, K., Baker, J.T., Stufflebeam, S.M., Wang, K., Wang, X., Hong, B., Liu, H., 2015. Parcellating cortical functional networks in individuals. Nat Neurosci 18, 1853–1860.

Wang, J., Zuo, X., He, Y., 2010. Graph-based network analysis of resting-state functional MRI. Fron Syst Neurosci 4, 16.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J Amer Statist Assoc 58, 236–244.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 393, 440–442.

Wig, G.S., Laumann, T.O., Petersen, S.E., 2014. An approach for parcellating human cortical areas using resting-state correlations. Neuroimage 93, 276–291.

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106, 1125–1165.

Zalesky, A., Fornito, A., Harding, I.H., Cocchi, L., Yücel, M., Pantelis, C., Bullmore, E.T., 2010. Whole-brain anatomical networks: Does the choice of nodes matter? Neuroimage 50, 970–983.

Zilles, K., Amunts, K., Brodmann, K., 2010. Centenary of Brodmann's map–conception and fate. Nat Rev Neurosci 11, 139–145.