Towards Effective Functional Connectome Fingerprinting

Kendrick Li and Gowtham Atluri(✉)

Department of EECS, University of Cincinnati, Cincinnati, OH 45221, USA
likt@mail.uc.edu, atlurigm@ucmail.uc.edu

Abstract. The ability to uniquely characterize individual subjects based on their functional connectome (FC) is a key requirement for progress towards *precision neuroscience*. The recent availability of dense scans from individuals has enabled the neuroscience community to investigate the possibility of individual characterization. FC fingerprinting is a new and emerging problem where the goal is to uniquely characterize individual subjects based on FC. Recent studies reported near 100% accuracy suggesting that unique characterization of individuals is an accomplished task. However, there are multiple key aspects of the problem that are yet to be investigated. Specifically, (i) the impact of the number of subjects on fingerprinting performance needs to be studied, (ii) the impact of granularity of parcellation used to construct FC needs to be quantified, (iii) approaches to separate subject-specific information from generic information in the FC are yet to be explored. In this study, we investigated these three directions using publicly available resting-state functional magnetic resonance imaging data from the Human Connectome Project. Our results suggest that fingerprinting performance deteriorates with increase in the number of subjects and with the decrease in the granularity of parcellation. We also found that FC profiles of a small number of regions at high granularity capture subject-specific information needed for effective fingerprinting.

Keywords: Functional connectivity · Fingerprinting · Parcellation
Precision neuroscience

1 Introduction

Resting state functional connectivity (RSFC) studies that estimate connectivity based on blood-oxygen-level-dependent (BOLD) signal measured using functional magnetic resonance imaging (fMRI) have revealed many principles of brain function [4,5]. Most of the existing studies made inferences about RSFC at a group level, by co-registering individual scans to a standard template, and found that such inferences are reliable [11]. While group-level inferences inform us of the generic principles, they obscure principles specific to individual subjects that are essential for characterizing brain function in health and disease. Recent availability of 'dense' fMRI scans from individuals (e.g., Human Connectome Project

G. Wu et al. (Eds.): CNI 2018, LNCS 11083, pp. 107–116, 2018.
https://doi.org/10.1007/978-3-030-00755-3_12

(HCP) data [12], Midnight Scan Club (MSC) [7], and MyConnectome dataset [8]) provide a tremendous opportunity to study idiosyncratic properties of brain function and make progress towards 'precision neuroscience' [10].

Functional connectome fingerprinting, where the goal is to identify individuals using subject-specific RSFC, has been explored using the above datasets that constitute dense scans from individuals [6,9]. Specifically, given a set of N *reference* fMRI scans, one from each of the N subjects, and a new *target* fMRI scan from one of the same N subjects, the goal is to identify the subject by 'matching' RSFC of the target scan with that of the reference scans. As RSFC is used to match the reference and the target scans, we refer to it as a functional fingerprint. There are different approaches to using RSFC and their effect on the accuracy of fingerprinting has been studied. For instance, Finn et al. [6], using 126 subjects from HCP, reported a fingerprinting accuracy in the range of 92%–94% while using whole-brain RSFC and 98–99% using a frontoparietal-based RSFC. In another study, using 100 unrelated subjects from HCP, Amico and Goni [3] observed that by performing principal component analysis (PCA) on whole-brain RSFC and using the resultant principal components for matching, the accuracy increased from 94% to 98%. Xu et al. [15] studied the reliability of boundaries drawn between functional areas delineated using spatial gradients (the approach is discussed elaborately in [14]) and reported success rate of up to 99% using 30 subjects.

These near 100% success rates may lead one to conclude that fingerprinting is not only a relatively easy problem, but also a solved problem with no room for progress. However, this is far from reality. Note that the underlying hypothesis that drives the fingerprinting methodology is that RSFC instances from the same subject lie in close proximity, segregated from other subjects' RSFC, in some high-dimensional space. When a small number of subjects are sampled from a population, the RSFCs from one subject may be well separated from that of others in the high-dimensional space. However, when many more subjects are sampled from a population, this high-dimensional space may become cluttered with RSFCs from different subjects, where RSFCs from different subjects may look more similar than the RSFCs from the same subject, and as a result hurt the overall fingerprinting performance. This aspect of fingerprinting is yet to be studied.

In addition, the impact of granularity of the parcellation used for computing RSFC on fingerprinting accuracy is yet to be investigated. A parcellation of the brain is expected to capture functionally distinct areas at a given level of granularity, often indicated as the number of parcels. While *subject-specific* RSFC is desired for fingerprinting purpose, the granularity at which this subject-specific information becomes available is not known.

Subject-level RSFC contains both generic and subject-specific information. Separating out subject-specific information from generic information is crucial for determining what aspects of RSFC are relevant for fingerprinting. Finn et al.'s approach [6] of using RSFC within different groups of brain regions is one approach. Their underlying hypothesis is that the subject-specific signatures are present within the RSFC of different region groups. One hypothesis, that is not yet explored, is that an RSFC profile of one or small number of regions could be used for fingerprinting. This direction allows us to study the degree of subject-specific information available in a single-node's FC profile and it also allows us to discover the regions in the brain that provide subject-specific connectivity maps for fingerprinting.

In this study, we investigated the above directions to deepen our understanding of functional connectome fingerprinting. Specifically, we addressed the following three questions: (1) How does the number of subjects affect the accuracy of fingerprinting? (2) How does the granularity of parcellation used for computing RSFC affect fingerprinting accuracy? (3) Can we find RSFC elements that are highly suited for effective fingerprinting? We performed our analysis on resting state fMRI data from 339 unrelated individuals in the HCP, using computing resources from the Ohio Supercomputer Center [16]. Our results suggest that fingerprinting performance deteriorates with increase in the number of subjects and with the decrease in the granularity of parcellation. We also found that a small number of regions at high granularity capture subject-specific information needed for effective fingerprinting.

The rest of this paper is organized as follows: The datasets used in our study are described in Sect. 2. Methods we used to answer above questions are presented in Sect. 3. We discussed our results in Sect. 4 and we concluded with Sect. 5.

2 Data

Resting state fMRI data from the 1200-subjects 2017 HCP data release (March 2017) [12] was used in this study. This release included processed resting state fMRI scans from 1003 healthy young adults. While we could use all of the 1003 subjects' data, any familial relationships among subjects may muddle our analysis for fingerprinting. To avoid familial relationships among subjects, we used a set of 339 unrelated subjects provided in the HCP release [2].

As part of the HCP, resting-state fMRI scans were collected from each subject on two separate days. On each day, a 20 min scan left-to-right (LR) phase encoded scan and a 20 min right-left (RL) phase encoded scan were obtained. For these four fMRI scans, we used the extensively-preprocessed node-timeseries data that was made available in the HCP data release. This node-timeseries data was generated by performing a series of steps including preprocessing, artefact removal using ICA, inter-subject registration, group-PCA, group-Independent Component Analysis (ICA), and dual-regression to compute time series for each independent component (IC). These steps are described in the HCP documentation [1]. As part of the Group-ICA step of the HCP preprocessing pipeline, the brain was parcellated into ICs at different granularities: 15, 25, 50, 100, 200, and 300 regions. Node-timeseries for ICs from each of these parcellations were provided in the HCP data release. We refer to the set of node timeseries from these ICs as IC_{15}, IC_{25}, IC_{50}, IC_{100}, IC_{200}, and IC_{300}. The node-timeseries data from the March 2017 release was used as is without further processing.

3 Methods

3.1 FC Fingerprinting

We will formally establish the terminology that will be used in the rest of the paper. We refer to fMRI scans for which we know which subject they are collected from as ‘reference’ scans. We refer to the new set of scans for which the subject they are collected from needs to be determined by matching with reference scans as ‘target’ scans. Given a set of N reference scans $\{R_1, R_2, \ldots, R_N\}$ from N different subjects, and a set of target scans $\{T_1, T_2, \ldots, T_N\}$ from the same set of subjects, the problem of FC fingerprinting is to determine for each target scan T_i the corresponding subject’s reference scan R_j by matching their RSFC. There are two key steps here: (1) computing FC, (2) matching FC.

For computing RSFC from an IC node-timeseries derived from a scan, we computed Pearson correlation between each pair of node-timeseries. As the scans were collected from each subject on two separate days, we computed the average RSFC per day. That is, we averaged the RSFC from the resting-state LR and RL encoding scans on each day. As a result we have two RSFCs, one per day, from each subject: RSFC_{d1} and RSFC_{d2}. As the node-timeseries data is available for different granularities of parcellation, these two RSFCs were computed for each of the granularities.

For matching RSFCs, we used a method that is similar to that of Finn et al.’s [6] whole-brain approach. Specifically, for each RSFC computed from a target scan T_i, we computed the Pearson correlation between the vector constructed by taking the upper-triangular values of the target RSFC matrix with that of each of the reference RSFCs. The reference RSFC that showed highest correlation with the target RSFC is treated as a match.

The accuracy of fingerprinting is computed as the fraction of subjects for which the target scans were perfectly matched with their reference scans. As we have two RSFCs from each subject (RSFC_{d1} and RSFC_{d2}), we computed fingerprinting accuracy in two ways: (1) using RSFC_{d1} as a reference and RSFC_{d2} as target, (2) using RSFC_{d2} as a reference and RSFC_{d1} as target. The results from the former and latter cases are labelled as *Day1 Ref: Day2 Tgt* and *Day 2 Ref: Day1 Tgt*, respectively.

3.2 Studying the Effect of Sample Size

To study the effect of sample size on fingerprinting accuracy, we conducted fingerprinting analysis on smaller subsets of the dataset. Out of the 339 subjects in our dataset, we randomly selected samples of different sizes ($\{50, 95, 140, 185, 230, 275, 320\}$) and computed their fingerprinting accuracies using the method described above. This was repeated 100 times for each size and the average accuracies for the 100 runs are reported.

Silhouette Coefficient Based Analysis: To investigate the effect of the sample size further and to test our hypothesis that with more and more subjects the RSFC space gets cluttered making it difficult to perform fingerprinting accurately, we used Silhouette coefficient [13], a commonly used cluster evaluation metric, to determine how well separated the subjects' RSFCs are in the space. A Silhouette value is computed for each data point in the cluster and the value can only range from -1 to 1. Positive values closer to 1 indicate that the data point is at the core of the cluster, while a value closer to -1 indicates that the point is actually closer to points in another cluster than in the same cluster. For our analysis, an RSFC with a negative value is indicative that it is more similar to RSFCs from other subjects than it is to the RSFCs from the same subject. For a more complete treatment of Silhouette coefficient we refer an interested reader to [13]. The Silhouette value was calculated for each RSFC by assigning all RSFCs from each subject to a separate cluster. For this analysis, we used RSFCs computed from all four scans of a subject and so each cluster has four members. We computed the average Silhouette value over all RSFCs from all subjects. To understand how sample size affects the space of RSFCs using Silhouette coefficient, we created 100 randomly sampled sets of subjects each for different sample-sizes ($\{5, 50, 95, 140, 185, 230, 275, 320\}$). For each sample-size, we computed the average Silhouette coefficient for each of the hundred sets. We also computed the fraction of subjects that contained a scan with a negative Silhouette value in each of the sets for different sample-sizes. We reported the average of the fraction of subjects.

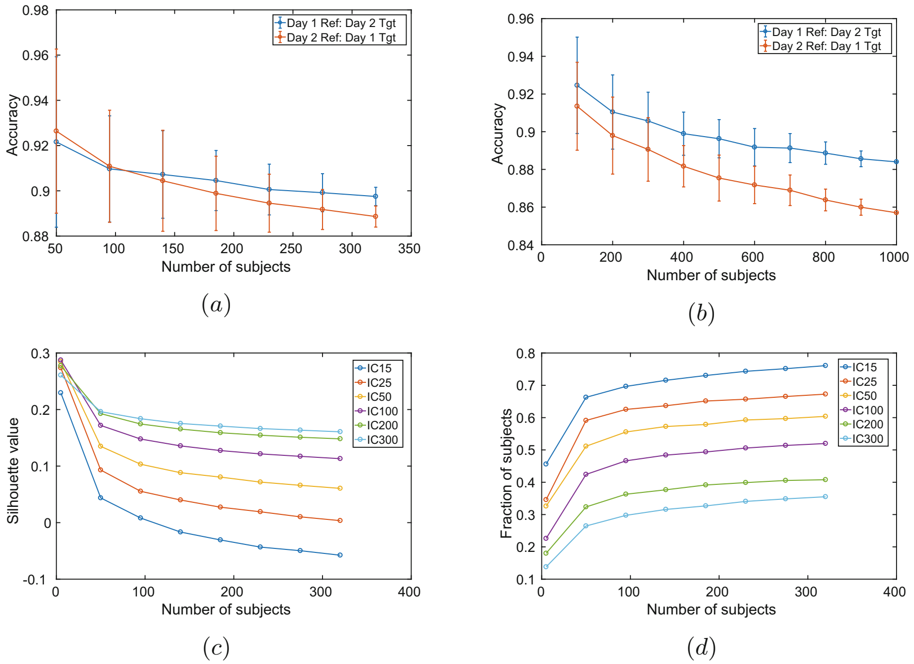
3.3 Studying the Effect of Granularity of Parcellation

To study the effect of the granularity of parcellation on fingerprinting accuracy, we performed fingerprinting analysis on 100 randomly sampled subjects using their node-timeseries from IC_{15}, IC_{25}, IC_{50}, IC_{100}, IC_{200}, and IC_{300}. For each level of granularity, the fingerprinting accuracy was recorded. This was repeated 100 times and the average accuracy for each granularity are reported.

3.4 Determining Elements of RSFC that are Highly Relevant for Fingerprinting

Subject-level RSFC contains both generic and subject-specific information. Separating out subject-specific information from generic information is crucial for determining what elements of RSFC are relevant for fingerprinting. We pursue two key methods for identifying relevant RSFC components. Our approach is to select the RSFC profile for one brain region, i.e., all region-pairs that involve the brain region, and compute the fingerprinting performance of that region's FC profile.

Single-Node RSFC Based Fingerprinting. The FC fingerprinting method described above (in Sect. 3.1) uses the entire set of elements from an RSFC. Our hypothesis is that only one region's connectivity profile may be sufficient to uniquely fingerprint a subject. To test this, we use only edges incident on one



(a)

(b)

(c)

(d)

Fig. 1. The effect of the number of subjects on RSFC fingerprinting. (a) Average accuracy of fingerprinting as the number of subjects increased from 50 to 320 using the IC_{300} dataset. The error bars indicate accuracies one standard deviation away from the mean. (b) Accuracy of fingerprinting on the 1000 subject IC_{300} dataset as the number of subjects increased from 100 to 1000. Error bars indicate accuracies one standard deviation away from the mean. (c) The average subject Silhouette values with varying number of subjects. (d) The fraction of subjects with a negative Silhouette value for at least one RSFC.

region (at a time) for matching a target RSFC with reference RSFCs. This is repeated for all the regions in the parcellation to determine the regions whose RSFC profile captures highly subject-specific information. We randomly selected 100 subjects and conducted the fingerprinting analysis for each parcellation granularity on each node and recorded the resultant accuracies.

Studying Reliability of Single Node Analysis. We randomly selected a set of 150 subjects from the 339 unrelated individuals and we refer to it as ‘Group A’. From the remaining individuals, we randomly selected another set of 150 subjects and refer to it as ‘Group B’. Using RSFCs computed from IC_{300} dataset, we computed for each of the 300 nodes their FC fingerprinting accuracy when single node RSFC is used for groups A and B separately. We compare these node-level fingerprinting accuracies from groups A and B to determine the reliability of the single node fingerprinting accuracies.

4 Results

4.1 The Effect of Sample Size on Fingerprinting

The results from our analysis of quantifying the effect of sample size on fingerprinting are shown in Fig. 1. The fingerprinting accuracy for unrelated individuals decreased from 92.16% to 89.75% as the number of subjects increased from 50 to 320 for the scenario *Day 1 Ref: Day 2 Tgt*. Similar reduction in accuracies were seen for *Day 2 Ref: Day 1 Tgt*, even though these accuracies are relatively small compared to *Day 1 Ref: Day 2 Tgt* (Fig. 1(a)). Note that the observed (2–3%) decrease in fingerprinting accuracy (in Fig. 1(a)) may seem tolerable, but when FC fingerprinting is considered for clinical practice where the underlying sample size is significantly larger the estimated accuracy may not meet the demands of precision neuroscience. To understand the extent of this drop in accuracy on larger datasets, we performed this analysis on the larger HCP dataset with 1000 subjects. The accuracy for 1000 subjects was 85.8%, for the scenario *Day 2 Ref: Day 1 Tgt*. These results suggest that there can be a significant reduction in accuracies as larger and larger datasets are considered.

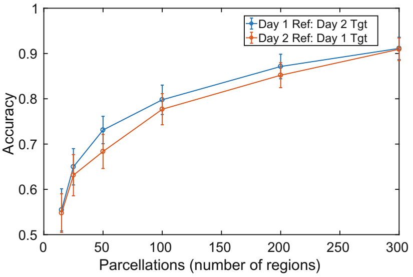


Fig. 2. The accuracy of fingerprinting with change in parcellation granularity. The error bars indicate accuracies one standard deviation away from the mean.

In general, this reduction in accuracy could be due to RSFCs from different subjects exhibiting more similarity than the RSFCs from the same subject with the increase in the number of subjects. That is, the space of RSFCs is more cluttered as the number of subjects increased. To further investigate this hypothesis of cluttering in RSFC space due to increased number of subjects, we used Silhouette coefficient, a popular cluster evaluation metric, to quantify segregation of RSFCs. The average subject Silhouette value decreased from 0.2608 to 0.1606 as the number of subjects increased from 5 to 320 subjects for IC_{300} (Fig. 1(c)). This supports our hypothesis that the space of RSFCs becomes less segregated (or more cluttered) as the number of subjects increased. Furthermore, there was an increase in the fraction of subjects with a negative Silhouette value for at least one RSFC from an average value of 13.8% to 35.51% as the number of subjects increased from 5 to 320 for IC_{300} (Fig. 1(d)). This quantifies the degree of cluttering in the RSFC space as a function of sample size.

4.2 The Effect of Parcellation Granularity on Fingerprinting

We also saw an increase in fingerprinting accuracy as the granularity of parcellation increased (Fig. 2). The average accuracy increased from 55.47% to 91.13% as the number of parcels increased from 15 to 300 for the scenario *Day 1 Ref:*

Day 2 Tgt. This suggests that finer parcellations capture subject-specific RSFC more effectively than coarser parcellations. This result is also in agreement with our previous Silhouette results (Fig. 1(c)); in all cases the Silhouette values were lower, and fraction of subjects with a negative Silhouette value were higher, when coarse parcellation was used (Fig. 1(c) and (d)).

We also performed a combined analysis on the effect of the number of subjects and granularity of parcellation. The average accuracy showed a constant downward trend with an increase in the number of subjects and a constant upward trend with an increase in the number of parcels (Fig. 3).

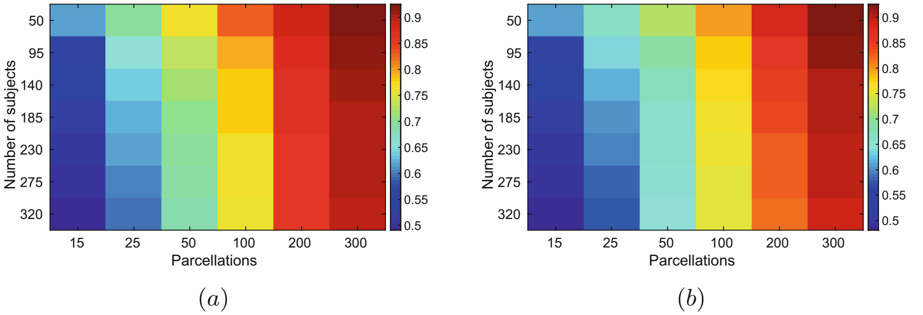


Fig. 3. Heatmap showing the relation between the number of subjects and the granularity of parcellation on fingerprinting accuracy. (a) *Day 1 Ref: Day 2 Tgt* (b) *Day 2 Ref: Day 1 Tgt*.

4.3 Determining Elements of RSFC that Are Highly Relevant for Fingerprinting

We computed fingerprinting accuracy for each brain region by 'matching' the edges incident on the region from the target RSFC with the reference RSFCs. This was repeated for each parcellation granularity. The results are shown in Fig. 4(a) and (b). There are three key observations: (1) There is an increase in the range of fingerprinting accuracy as the number of nodes increased (Fig. 4(a) and (b)). (2) The best single-node accuracy for finer parcellations are nearly as good as the whole-brain RSFC based accuracy. For instance, best single-node accuracy for IC_{300} was 86.13% compared to the whole-brain RSFC accuracy of 91.13% (only 5% lower) for *Day 1 Ref: Day 2 Tgt*. (3) The difference between the best single-node accuracy and the whole-brain RSFC decreased with increase in the number of parcellations. These results suggest that at a finer granularity of parcellation, some region's RSFC not only captures subject-specific information but also does so nearly as well as the whole-brain RSFC.

To assess the reliability of the accuracies across two different samples of subjects, we created two non-overlapping groups A and B of 150 subjects each and computed single node RSFC based accuracies separately. The single-node

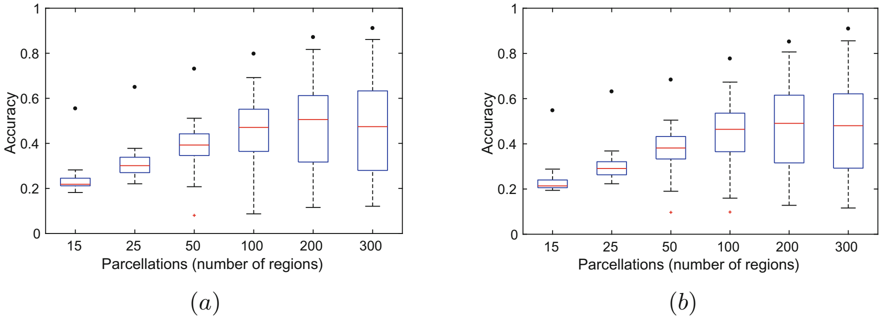


Fig. 4. Single node RSFC based fingerprinting accuracy: (a) *Day 1 Ref: Day 2 Tgt* (b) *Day 2 Ref: Day 1 Tgt.* The accuracy of using the full RSFC for fingerprinting as a black dot for each parcellation granularity.

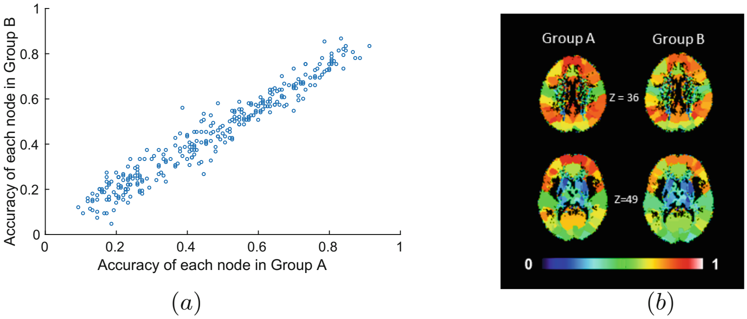


Fig. 5. (a) Comparision between the single-node RSFC-based fingerprinting accuracy between groups A and B for *Day 1 Ref: Day 2 Tgt.* (b) The accuracies of each node are colored in the brain volume for groups A and B.

accuracies of 300 nodes in the IC_{300} parcellation are strongly correlated between groups A and B (Fig. 5(a)). This suggests that these regions that consistently resulted in higher accuracies in independent samples capture FC information unique to individual subjects. The fingerprinting accuracies for the components in the IC_{300} dataset for groups A and B are shown in Fig. 5(b). The regions that resulted in higher accuracies in group A also resulted in higher accuracies in group B. Particularly, the ICs in the frontal region and lateral-parietal regions resulted in highest accuracy, approximately 90%, among other regions. These results are consistent with the findings reported in Finn et al. [6], where they observed frontoparietal network to exhibit very high accuracy.

5 Conclusion

In this work we investigated the different aspects of FC fingerprinting that have been overlooked. They include the impact of number of subjects and granularity of parcellation. We also studied single-node RSFC-based fingerprinting and the reliability of the resultant accuracies. Our results suggest that as the number of subjects increase the RSFC space gets more and more cluttered resulting in reduced accuracies. We borrowed ideas from cluster evaluation that have been well studied in the data mining community. We also found that with a high-granularity of parcellation, higher fingerprinting accuracies are possible. We also investigated the role of single-node RSFC in effective fingerprinting. We found that just one brain region's RSFC profile can be nearly as good as the whole-brain RSFC based matching. We also observed that the frontal and lateral-parietal regions that show very high accuracies are also reliable across independent samples.

References

1. HCP documentation. https://www.humanconnectome.org/storage/app/media/documentation/s1200/HCP1200-DenseConnectome+PTN+Appendix-July2017.pdf
2. List of 339 unrelated individuals in HCP. https://wiki.humanconnectome.org/download/attachments/89391484/Unrelated_S900_Subject_multilist1_with_physio.csv
3. Amico, E., Goñi, J.: The quest for identifiability in human functional connectomes. Sci. Rep. **8**(1), 8254 (2018)
4. Atluri, G., et al.: The brain-network paradigm: using functional imaging data to study how the brain works. IEEE Computer **49**(10), 65–71 (2016)
5. Bandettini, P.A.: Twenty years of functional MRI: the science and the stories. Neuroimage **62**(2), 575–588 (2012)
6. Finn, E.S., et al.: Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. **18**(11), 1664 (2015)
7. Gordon, E.M., et al.: Precision functional mapping of individual human brains. Neuron **95**(4), 791–807 (2017)
8. Laumann, T.O., et al.: Functional system and areal organization of a highly sampled individual human brain. Neuron **87**(3), 657–670 (2015)
9. Miranda-Dominguez, O., et al.: Connectotyping: model based fingerprinting of the functional connectome. PloS one **9**(11), e111048 (2014)
10. Poldrack, R.A.: Precision neuroscience: dense sampling of individual brains. Neuron **95**(4), 727–729 (2017)
11. Shehzad, Z., et al.: The resting brain: unconstrained yet reliable. C.Cortex (2009)
12. Smith, S.M.: Resting-state fMRI in the human connectome project. Neuroimage **80**, 144–168 (2013)
13. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining (2006)
14. Wig, G.S.: An approach for parcellating human cortical areas using resting-state correlations. Neuroimage **93**, 276–291 (2014)
15. Xu, T.: Assessing variations in areal organization for the intrinsic brain: from fingerprints to reliability. Cereb. Cortex **26**(11), 4192–4211 (2016)
16. Ohio Supercomputer Center (1987). http://osc.edu/ark:/19495/f5s1ph73