Maximizing the individual fingerprints of human functional connectomes through decomposition into brain connectivity modes.

Enrico Amico [1,2] and Joaquín Goñi [1,2,3,*]
[1] School of Industrial Engineering, Purdue University, West-Lafayette, IN, USA
[2] Purdue Institute for Integrative Neuroscience, Purdue University, West-Lafayette, IN, USA
[3] Weldon School of Biomedical Engineering, Purdue University, West-Lafayette, IN, USA
[*] Correspondence: jgonicor@purdue.edu

Abstract

The evaluation of the individual “fingerprint” of a human functional connectome (FC) is becoming a promising avenue for neuroscientific research, due to its enormous potential inherent to drawing single subject inferences from functional connectivity profiles. Here we show that the individual fingerprint of a human functional connectome can be maximized from a reconstruction procedure based on group-wise decomposition in a finite number of brain connectivity modes. We use data from the Human Connectome Project to demonstrate that the optimal reconstruction of the individual FCs through connectivity eigenmodes maximizes subject identifiability across resting-state and all seven tasks evaluated. The identifiability of the optimally reconstructed individual connectivity profiles increases both at the global and edgewise level, also when the reconstruction is imposed on additional functional data of the subjects. We extend this approach to also map the most task-sensitive functional connections. Results show that is possible to maximize individual fingerprinting in the functional connectivity domain regardless of the task, a crucial next step in the area of brain connectivity towards individualized connectomics.

Introduction

The explosion of publicly available neuroimaging datasets in the last years have provided an ideal benchmark for mapping functional and structural connections in the human brain. At the same time, quantitative analysis of connectivity patterns based on network science have become more commonly used to study the brain as a network[1] , giving rise to the area of research so called Brain Connectomics[2,3]. The analyses of functional and structural brain connectivity patterns have allowed researchers to make inferences on the different organization of brain networks in clinical and healthy populations, and to identify changes in these cohorts, usually by testing differences across groups [4,5]. Until recently, brain connectivity studies have generally overlooked the existing connectivity heterogeneity within each group, for several reasons. The group average procedure eases comparisons between different populations and has the benefit of providing more representative connectivity patterns. However, this comes at the cost of ignoring the potentially precious information provided by subject level, i.e. individual , connectomes.

The recent work by Finn et al. [6] on fingerprinting has paved the way to the new promising avenue of detecting individual differences through brain connectivity features. They showed that the individual functional connectivity (FC) profiles estimated from functional resonance

magnetic imaging (fMRI) data can be seen as a “fingerprint” of the subject, which indeed may be used to identify a given individual in a set of functional connectivity profiles from a population.

The capacity of functional connectivity profiles to identify subjects goes along with the concept of reproducibility of test-retest experiments [6]. The rationale behind is that the higher the accuracy of the functional connectivity on each fMRI session and subject, the better will be the identifiability of individual subjects. Several aspects may have an impact on the quality, and hence on the identifiability of the data. This includes the characteristics of the fMRI sequence (such as its spatial and temporal resolution [7]), the processing of the fMRI data (including how to handle head motion and other artifacts [8,9] and the statistical approach used to obtain pairwise region-to-region functional connectivity from voxel-level time-series [4,10].

All the aspects listed above refer to efforts on increasing the accuracy of functional connectivity on individual fMRI sessions. Indeed, all of them could be applied by acquiring just one fMRI session of one subject. In the lack of gold-standards in brain connectivity, it is important to investigate the accuracy of connectome fingerprinting by procedures that gradually assess the data from cohort to individuals.

If individual fingerprinting based on functionally connectivity could be presented as a continuum, one could think of an approach that gradually goes from common connectivity patterns highly present in the cohort to individual patterns present in certain subjects or even in only certain individual sessions. Such a framework would allow us to decipher between what connectivity patterns are common in a cohort, what connectivity traits are unique of different individuals and what are session-dependent or beyond, i.e. spurious patterns from the standpoint of individual fingerprinting.

Here we propose a group-level framework to assess and maximize the individual fingerprinting of functional connectomes based on a principal component analysis (PCA) decomposition and subsequent individual reconstruction. We show that the uniqueness of each individual connectivity profile can be reconstructed through an optimal finite linear combination of orthogonal principal components (or eigenmodes) in the connectivity domain, hence here denominated brain connectivity modes. These connectivity modes improve the identification of each individual’s functional architecture both at the whole-brain and local sub-network level. We evaluate this methodology on 100 unrelated subjects of the Human Connectome Project (HCP), for test-retest data including resting-state and 7 different task-fMRI (see Methods).

The impact of the decomposition into connectivity modes and subsequent reconstruction of FC patterns is assessed in different scenarios. For all 7 fMRI tasks and resting-state, we find the existence of optimal reconstructions that maximize individual identifiability of functional connectomes. At those optimal solutions, edgewise identifiability as measured by intra-class-correlation is largely enhanced. The possible influence of motion (absolute frame displacement per session) in individual identifiability is assessed and found to be significant for all fMRI tasks but not for resting-state. We propose a generalization of this framework for cross-sectional data based on splitting the fMRI time-series into two halves and evaluate it on all four resting-state sessions. Finally, we map task-specific functional edges as measured by intra-class correlation and identify which within and between resting-state-networks (RSNs) are the most task-specific.

We conclude by discussing the interpretation of the concept of brain connectivity modes, or eigenmodes in the functional connectivity domain. We make considerations on possible

driving factors (mainly motion and task performance) that may limit the maximization of individual identifiability. Finally, we discuss the limitations of our study and future work and potential applications of this methodology.

Materials and Methods

Dataset. The fMRI dataset used in this work is from the Human Connectome Project (HCP, http://www.humanconnectome.org/), Release Q3. Below is the full description the acquisition protocol and processing steps.

HCP data. We used the 100 unrelated subjects from the HCP 900 subjects data release[10]. The fMRI resting-state runs (HCP filenames: rfMRI_REST1 and rfMRI_REST2) were acquired in separate sessions on two different days, with two different acquisitions (left to right or LR and right to left or RL) per day [10,11]. The seven fMRI tasks were the following: gambling (tfMRI_GAMBLING), relational (tfMRI_RELATIONAL), social (tfMRI_SOCIAL), working memory (tfMRI_WM), motor (tfMRI_MOTOR), language (tfMRI_LANGUAGE, including both a story-listening and arithmetic task) and emotion (tfMRI_EMOTION). The working memory, gambling and motor task were acquired on the first day, and the other tasks were acquired on the second day[11,13]. The HCP scanning protocol was approved by the local Institutional Review Board at Washington University in St. Louis. For all sessions, data from both the left-right (LR) and right-left (RL) phase-encoding runs were used to calculate connectivity matrices. Full details on the HCP dataset have been published previously [11,12,14].

Brain atlas. We employed a cortical parcellation into 360 brain regions as recently proposed by Glasser et al. [15]. For completeness, 14 sub-cortical regions were added, as provided by the HCP release (filename “Atlas_ROI2.nii.gz”). To do so, this file was converted from NIFTI to CIFTI format by using the HCP workbench software [14,15](command *-cifti-create-label* http://www.humanconnectome.org/software/connectome-workbench.html)

HCP preprocessing: functional data. The HCP functional preprocessing pipeline [11,12] was used for the employed dataset. This pipeline included artifact removal, motion correction and registration to standard space. Full details on the pipeline can be found in [12,14]. The main steps were: spatial (“minimal”) pre-processing, in both volumetric and grayordinate forms (i.e., where brain locations are stored as surface vertices [12]); weak highpass temporal filtering (> 2000s full width at half maximum) applied to both forms, achieving slow drift removal. MELODIC ICA [17] applied to volumetric data; artifact components identified using FIX [18]. Artifacts and motion-related time courses were regressed out (i.e. the 6 rigid-body parameter time-series, their backwards-looking temporal derivatives, plus all 12 resulting regressors squared) of both volumetric and grayordinate data[11].
For the resting-state fMRI data, we also added the following steps: global gray matter signal was regressed out of the voxel time courses [9]; a bandpass first-order Butterworth filter in forward and reverse directions [0.001 Hz, 0.08 Hz] [9] was applied (Matlab functions *butter* and *filtfilt*); the voxel time courses were z-scored and then averaged per brain region, excluding outlier time points outside of 3 standard deviation from the mean, using the workbench software [16] (workbench command *-cifti-parcellate*). For task fMRI data, we applied the same

above mentioned steps except the bandpass filter, since it is still unclear the connection between different tasks and optimal frequency ranges [19].
Pearson correlation coefficients between pairs of nodal time courses were calculated (MATLAB command *corr*), resulting in a symmetric connectivity matrix for each fMRI session of each subject. Functional connectivity matrices were kept in its signed weighted form, hence neither thresholded nor binarized.

Finally, the resulting individual functional connectivity matrices were ordered (rows and columns) according to 7 resting-state cortical sub-networks (RSNs) as proposed by Yeo and colleagues [20]. For completeness, an 8th sub-network including the 14 HCP sub-cortical regions was added (as analogously done in recent paper [21]).

PCA reconstruction of the individual connectivity profiles

Principal component analysis (PCA) is a statistical procedure [20] that transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation), ranked in descending order of explained variance of the initial data. PCA has been widely used for exploratory analysis of the underlying structure of data in many areas of science, from pattern recognition in genetics[24] to denoising and compression in image processing [25].

In this study we explored the use of PCA (MATLAB command *pca*) in the connectivity domain for improving the individual fingerprint in functional connectomes from a group-level perspective. The procedure starts by matching the number of principal components included with the number of functional connectomes of the dataset. By definition, this PCA-based decomposition accounts for 100% of the variance in the data. As mentioned above, components are ranked according to explained variance in descending order. The next step consists of the reconstruction of the individual functional connectomes as a function of the number of components included (see methodological scheme at Fig. 1). The rationale behind this analysis is that high-variance components might carry cohort-level functional connectivity information, whereas lower-variance components might carry subject-level functional connectivity information and, finally, lowest-variance components might carry noisy or artifactual connectivity information. By doing this iterative fine-grained exploration of number of components used, we are able to identify in a data-driven fashion what are the boundaries for group-level, individual-level, and artifactual-level components. That is, once extracted the main connectivity-based principal components (PCs), each individual connectivity profile is reconstructed based on its mean and the linear combination of the chosen PCs (see Fig. 1).

For 16 fMRI sessions (Resting-state and 7 fMRI-tasks, with test-retest for each), we explored the property of individual fingerprinting for different levels of reconstruction based on the number of ranked components used. In the next section we will define the function employed for evaluating the level of individual fingerprint at any reconstruction level.

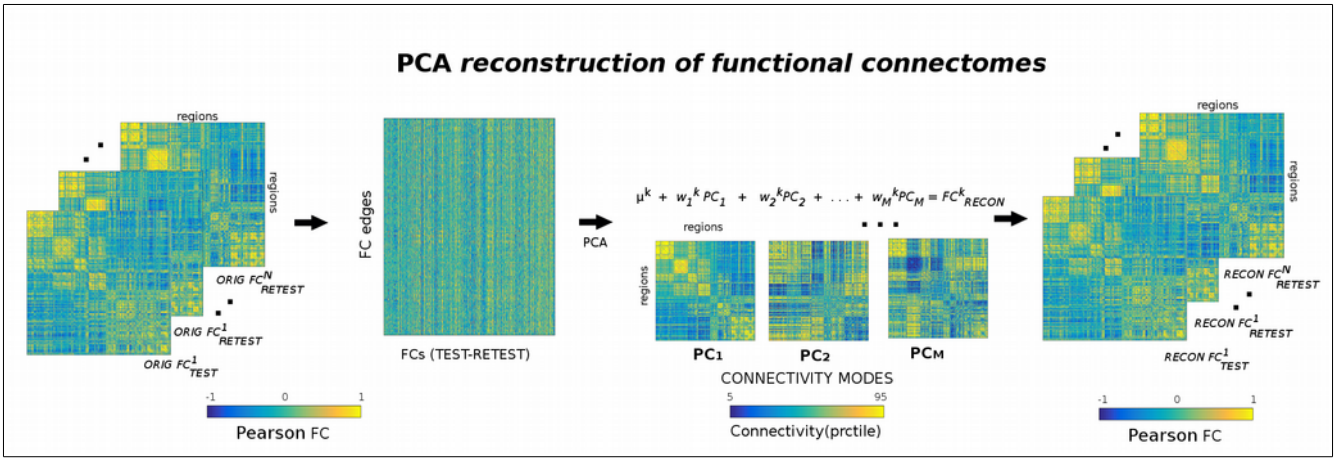


Figure 1. Workflow scheme of the group-level principal component analysis (PCA) reconstruction procedure of individual functional connectomes (FC). The upper triangular of each functional connectivity matrix (two FCs per subject, test-retest) is vectorized and added to a matrix where columns are sessions and rows are their vectorized functional connectivity patterns. Data are first centered: this is obtained by subtracting the mean μ^k from each column (where k goes from 1 to N subjects). Second, the PCA algorithm extracts the M principal components (i.e. the functional connectivity modes) associated to the whole population and their relative weights across subjects. The M orthogonal connectivity modes are then used to reconstruct back the FC of each subject (μ^k is added back to the data). Colorbars indicate positive (yellow) to negative (blue) connectivity values: Pearson's correlation coefficient in the case of individual FC matrices (left and right sides of scheme), and unitless connectivity weights in the case of PCA FC-modes. For ease of visualization, the FC-modes colorbar ranges from 5th to 95th percentile of the distribution of values.

Individual identifiability quality function

The maximization of a functional connectome fingerprint relies on the assumption that the connectivity profiles should be, overall, more similar between visits or sessions of the same subject than between different subjects. Finn et al. [6] showed that, to a great extent, it is possible to robustly identify the functional connectome of a subject "target" from a sample database of FCs, simply by computing the spatial (Pearson) correlation of the target FC against the database ones. For identification, they used a set of connectivity matrices from one session for the database and connectivity matrices from a second session acquired on a different day as the target set. Then, given a query connectivity matrix from the target set, they computed the correlations between this matrix and all the connectivity matrices in the database. Finally, for each query, the predicted identity was picked as the one with the highest correlation coefficient, and assigned a score of 1 if the predicted identity matched the true identity, and a score of 0 otherwise. The success rate of this identification procedure was above 90% for resting-state sessions, and ranged between 54% and 87% when including task-task and task-rest sessions [6].

In order to evaluate our framework and the identifiability capability as a continuum, we generalized the above mentioned binary score system [6] to a more continuous score on the level of individual fingerprinting present on a set of test-retest functional connectomes. Hence we introduce the concept of level of identifiability (I) on a set of functional connectomes.

Let **A** be the "identifiability matrix", i.e. the matrix of correlations (square, non symmetric) between the subjects' FCs from visit 1, and the FCs from visit 2. The dimension of **A** is N^2, where N is the number of subjects in the database. Let $\langle a_{ii} \rangle$ represent the average of the main diagonal elements of **A,** which consist of the Pearson correlation values between visits

of same subjects: from now on, we will refer to this quantity as “self identifiability”. Similarly, let $\langle a_{ij} \rangle$ define the average of the off-diagonal elements of matrix A, i.e. the correlation between visits of different subjects. Then we define the individual identifiability (**I**) of the population as the difference between both terms:

$$I = \langle a_{ii} \rangle - \langle a_{ij} \rangle \ , \ i \neq j$$

which quantifies the difference between the average within-subject FCs similarity and the average between-subjects FCs similarity. Since in our data both terms $\langle a_{ii} \rangle$ and $\langle a_{ij} \rangle$ were positive for all tasks, one can also define the percentage difference [26] of **I, namely I_{diff}:**

$$I_{diff} = \frac{\langle a_{ii} \rangle - \langle a_{ij} \rangle}{\frac{1}{2} * (\langle a_{ii} \rangle + \langle a_{ij} \rangle)} * 100 \ , \ i \neq j$$

I_{diff} is the percentage difference (range between 0 and 200) between the average correlation of two different visits of the same subjects and the average correlation of two different visits of different subjects. The higher the value of I_{diff}, the higher the individual fingerprint.

By defining I and subsequently I_{diff}, the optimization problem of individual identifiability is then reduced to maximizing I_{diff}. This consists of exploring within a range of number of components (M), and of finding the optimal number of components, m*, in the PCA decomposition that provides the maximum value of I_{diff}, namely I_{diff}*, for which:

$$I_{diff}^{*} = \arg\max_{m \in M} \ I_{diff}(m)$$

Edgewise individual identifiability and edgewise task identifiability

The function defined earlier provides a way to quantify the individual identifiability at a whole-network level. We also quantified the edgewise identifiability by using intraclass correlation [27,28] (ICC). ICC is a widely used measure in statistics, normally to assess the percent of agreement between units (or ratings/scores) of different groups (or raters/judges)[29]. It describes how strongly units in the same group resemble each other. The stronger the agreement, the higher its ICC value. We used ICC to quantify to which extent the connectivity value of an edge (functional connectivity value between two brain regions) could separate within and between subjects. In other words, the higher the ICC, the higher the individual identifiability of the connectivity edge.

Following the same rationale, one can also compute edgewise ICC when tasks are “raters” and “scores” given by subjects. In this case, the higher the ICC, the more separable the

different tasks across subjects and consequently the higher the task identifiability of the connectivity edge.

Influence of task-performance and motion regressors on self identifiability

Finally, we tested if optimal self identifiability (i.e. same subject's FC similarity) after PCA reconstruction was linked to task performance variables and/or motion estimators. HCP data collection provides an estimate of average absolute (i.e. displacement from initial frame, file name "Movement_AbsoluteRMS_mean.txt") and frame-to-frame displacement (file name "Movement_RelativeRMS_mean.txt") for each run and fMRI session. For each task session, response time and performance accuracy are also provided (file pattern {TASK_name}_stats.txt). We evaluated linear and $\log_{10}$-linear trends between self identifiability values after reconstruction and subject's motion displacement values (maximum between the two runs) and subject's task response time and accuracy (maximum, minimum and difference between the two runs)

Results

The dataset used for this study consisted of functional data from the 100 unrelated subjects in the Q3 release of the HCP [9,10]. For each subject, we estimated 18 functional connectivity matrices: 4 corresponding to resting-state (per HCP convention, named REST1_LR, REST2_LR, REST2_RL, REST1_RL,), 14 corresponding to each of the 7 tasks (test-retest, i.e. {TASK_NAME}_LR, {TASK_NAME}_RL, see Methods). The multimodal parcellation used here, as proposed by Glasser et al. [13,14], includes 360 cortical brain regions. We added 14 subcortical regions, hence producing functional connectome matrices (square, symmetric) of 374 x 374 (see Methods for details).

For each task (including resting-state), individual functional connectomes (including two visits, test-retest, per subject) were reconstructed based on PCA by iteratively including different number of components. This procedure can be summarized as follows (Fig. 1): first, the upper triangular part of each individual functional connectivity matrix was vectorized and added to a matrix where columns are the subjects and rows are their full connectivity pattern; second, the PCA algorithm extracted the principal components (PCs) associated to the whole population; third, these components were projected back in the individual subjects' space, leading to a "reconstructed" version of each original connectivity profiles (Fig. 1).

From the test-retest pool of 100 unrelated subjects (total of 200 FC matrices per task and 200 per resting-state, for this experiment only the REST1_LR and REST2_LR sessions were considered), a bootstrap technique was used to accurately estimate I_{diff} for each value m of number of components, m={2,5,10:10:160}. This was meant to avoid results driven by a small subset of the population. To do so, 100 random samples comprising the test-retest pairs of 80 subjects (total of 160 FC matrices) were performed for each value of m.

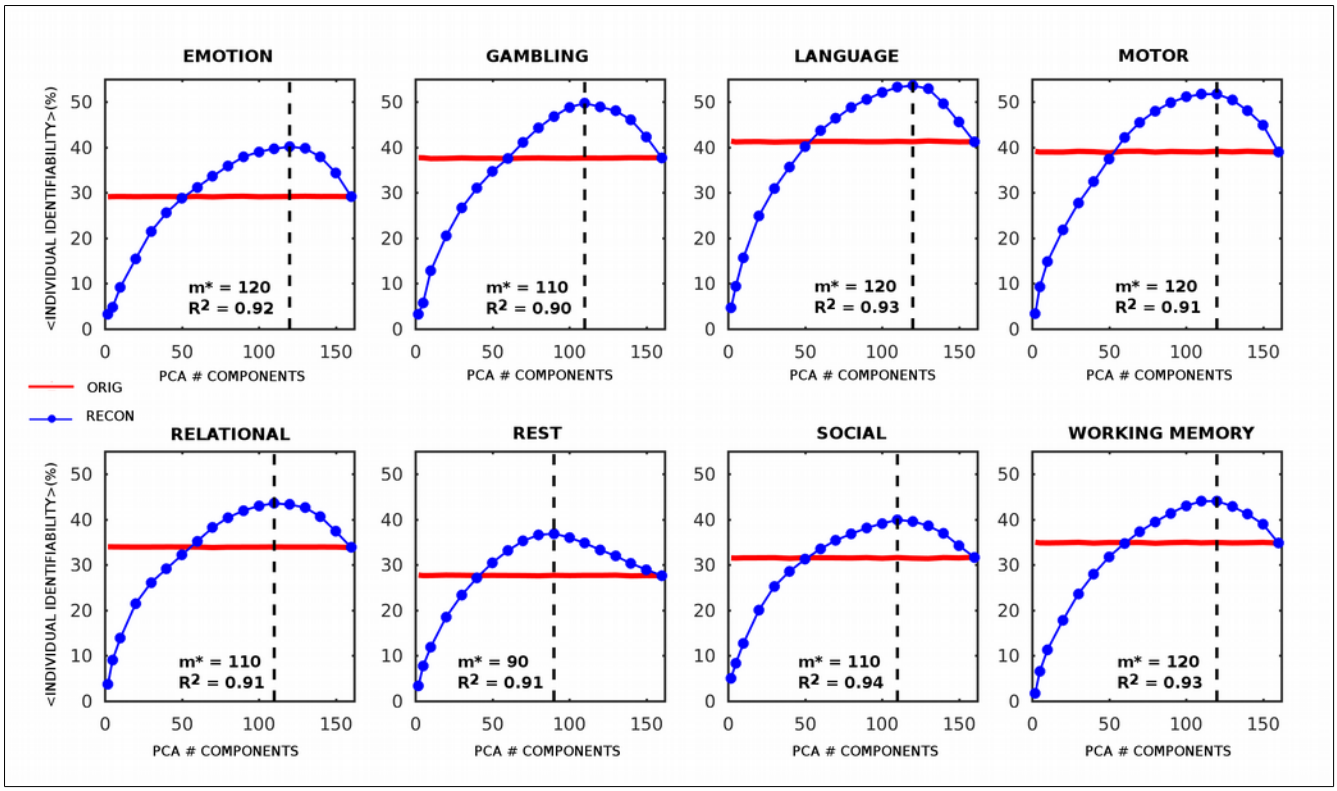


Figure 2. Percent difference of the individual identifiability (I_{diff}) as a function of the number of PCA components used for reconstruction in resting-state and 7 fMRI tasks. Plots show, for each task, the normalized individual identifiability as a function of the number of PCA components used for reconstruction (evaluated at 2, 5, and 10 to 160 components in steps of 10). Red line denotes the individual identifiability for the original FCs, whereas blue line with circles denotes the identifiability for reconstructed FCs based on the different number of components sampled. For each subplot, the optimal number of components that maximizes normalized individual identifiability (m*) and the corresponding explained variance (R2) are shown. The PCA reconstruction was tested on the resting state (REST) and 7 different task sessions provided by the HCP data (EMOTION, GAMBLING, LANGUAGE, MOTOR, RELATIONAL, SOCIAL, WORKING MEMORY, see Methods for details). To test the stability of the method, the individual identifiability was evaluated over 100 different runs. At each run, 80 subjects were randomly sampled from the HCP data pool of 100 unrelated subjects, 2 sessions (REST1_LR and REST2_LR for REST, {TASK_NAME}_LR and {TASK_NAME}_RL for the other 7 tasks, for a total of 160 FCs at every run). The standard deviation of I_{diff} (not shown in the plots) across runs was always lower than 0.8 %, for all the sessions considered, for both original and reconstructed data.

FC individual identifiability increases at optimal PCA reconstruction

For each number m of components retained for reconstruction, we tested the individual identifiability (I_{diff}) (see Methods) of the reconstructed connectomes with respect to the original ones, based on the following assumption: the connectivity profiles should be, overall, more similar between visits of the same subjects than between different subjects. For simplicity, in the main text we will show results for resting-state (REST) and for the MOTOR task . Results for all the other six tasks are provided in the supplementary material (supplementary Fig. 1 and supplementary Fig. 2). For all seven tasks and for resting-state, the PCA reconstructed functional connectomes outperformed the original ones in terms of I_{diff} for a wide range of m (Fig. 2, supplementary Fig. 1 and supplementary Fig. 2). Each

condition shows a slightly different optimal m* for maximizing I_{diff} (e.g. m*=90 PCs for resting-state, m*=110 for the MOTOR task, Fig. 2). For all cases, the variance of the functional data kept in the reconstruction was between 90% and 94%.

By definition, the individual identifiability optimization is constrained to the availability of test-retest fMRI sessions, which is not common in many acquired fMRI datasets, especially in clinical populations. To cover the necessity of assessing individual fingerprinting in these scenarios, we computed “two-halves” individual FCs by splitting each of the 4 resting state sessions in two parts (~600 fMRI volumes each). We then evaluated I_{diff} before and after reconstruction, with “test-retest” sessions now being the first and the second part of the same fMRI acquisition. Again, the PCA reconstructed functional connectomes keep outperforming the original ones for a wide range of m (see supplementary Fig. 3).

We next explored further properties of the optimal PCA-reconstruction for REST and the MOTOR task.

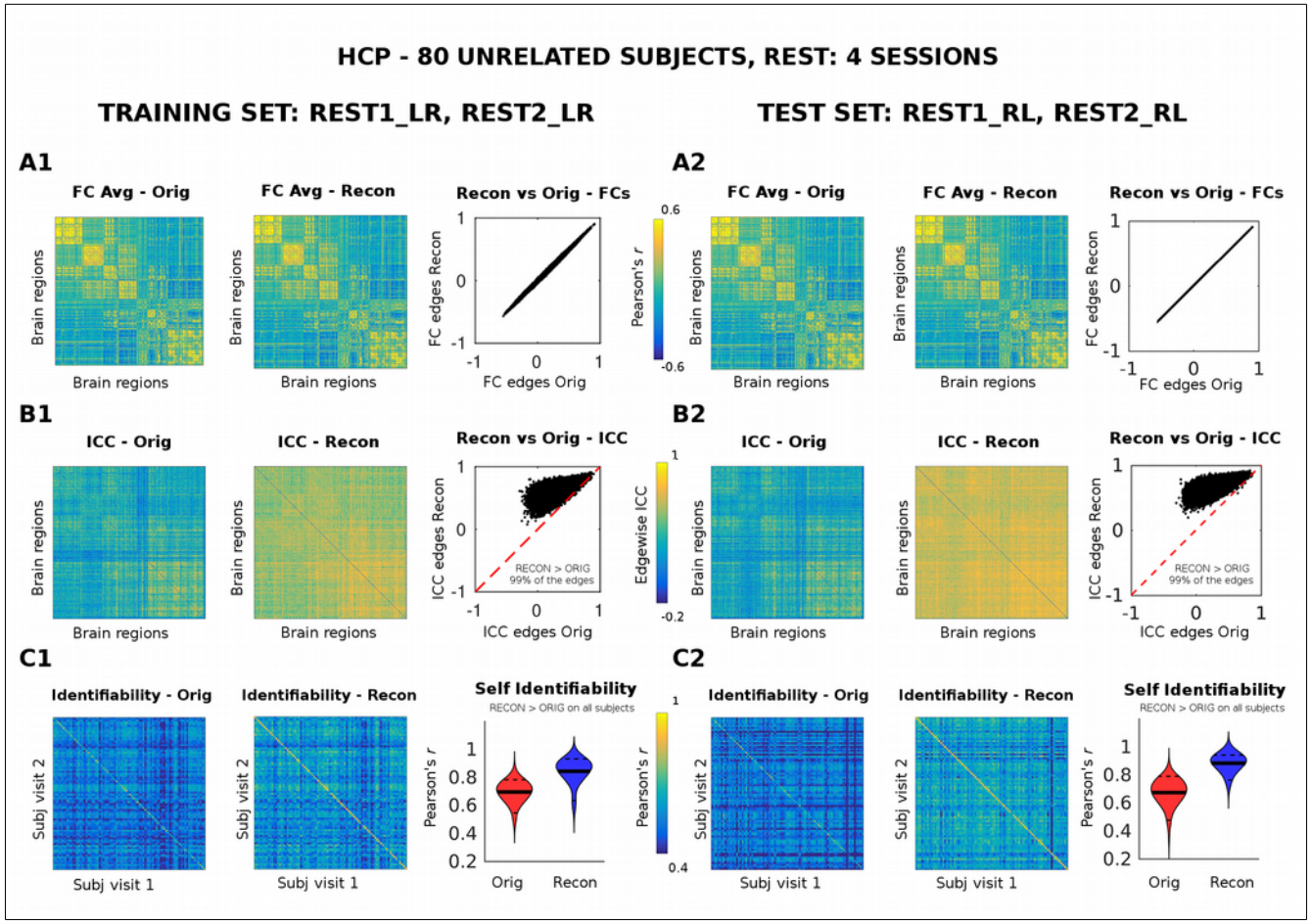


Figure 3. Evaluation and validation of PCA-reconstruction on resting-state functional connectomes (FCs) at the optimal point (m* = 90). The resting FCs of 80 subjects were reconstructed by using group-level PCA. The optimal number of PCA components was 90 (see Fig. 2). Results shown correspond to a single run. The PCA reconstruction was first evaluated on the REST1_LR and REST2_LR sessions (training set, left panel). The FC-modes extracted from the training set were then used to reconstruct two different resting state sessions REST1_RL and REST2_RL(test set, right panel). **A1-A2)** From left to right: the group averaged FC of the original data; the group averaged FC of the reconstructed data; the scatter plot edge by edge of the reconstructed group

averaged FC (y axis) vs original group averaged FC (x axis). **B1-B2)** From left to right: the intra-class correlation (ICC), computed over each FC edge, for the original data; the edgewise ICC for the reconstructed data; the scatter plot edge by edge of the reconstructed ICC values (y axis) vs original ICC values (x axis). The inset reports the percentage of edges where ICC increased after reconstruction (black dots on top of the red dashed line) from those that did not (black dots below of the red line). **C1-C2)** From left to right: Identifiability matrix (i.e. Pearson's correlation coefficient between functional profiles across subjects and sessions, see Methods) of the original data; identifiability matrix of the reconstructed data; violin plot of the "self identifiability" (i.e. the main diagonal of the identifiability matrix) distribution across the 80 subjects, for original (Orig, red) and reconstructed (Recon, blue). The solid black lines depict the mean value of the distribution; the dashed black lines the 5 and 95 percentiles. Note that, as specified by the inset, the self identifiability of each subject always improves after PCA reconstruction, both for the training and test set.

Reconstructed Functional connectomes: resting state

For the REST sessions, we tested the goodness of this method at the optimal PCA reconstruction point with respect to the original FCs in three different ways: 1) by comparing the group average functional connectomes before and after optimal PCA reconstruction; 2) by computing the edgewise intraclass correlation (ICC, i.e. how good a single edge can separate different subjects, see Methods), before and after optimal PCA reconstruction; 3) by evaluating the level of individual identifiability (as described in Methods), before and after optimal PCA reconstruction.

The 4 resting-state fMRI acquisitions per HCP subject available (i.e., REST1_LR, REST1_RL, REST2_LR, REST2_RL) allowed us to perform evaluation and validation of our methodology as follows: the optimal PCA reconstruction was first evaluated on the "training set" consisting of the REST_1 LR and REST2_LR sessions (Fig. 3, left panel), at the optimal number of 90 PCA components (as depicted in Fig. 2). The 90 FC-modes extracted from the training set were then used as the "orthogonal connectivity basis" through which reconstruct the two other FCs of the same subjects, i.e sessions REST1_RL and REST2_RL (Fig. 3, right panel).

Both for training and test sets, the optimal PCA reconstruction preserves the main characteristics of the functional connectomes (the group averaged functional connectomes before and after are almost identical, Fig. 3 A1,A2). Nonetheless, the edgewise ICC largely increase after optimal reconstruction for almost all edges (99%), both in the training and test cases (Fig. 3 B1, B2). In accordance with results shown in Fig.2, the self identifiability of the subjects' FCs after reconstruction increases on all subjects, regardless of whether one considers the LR or the RL acquisitions (Fig. 3 C1, C2).

Reconstructed Functional connectomes: motor task

Evaluation of the proposed methodology was also performed for all seven tasks available in the HCP dataset. We show as example results for the MOTOR task at the optimal reconstruction (m*= 110, Fig. 4). Also, supplementary Fig. 2 in the Supplementary material summarizes results for all the other tasks. Even in this case the reconstructed group average FC resembles the original one almost perfectly (Fig. 3A); edgewise ICC improves after optimal reconstruction on 89% of the functional edges (Fig. 3B); self identifiability increases after reconstruction on all subjects (Fig. 3C), with some subjects showing a larger increase than others (as shown by the violin plot distributions of self identifiability after reconstruction, Fig. 3C).

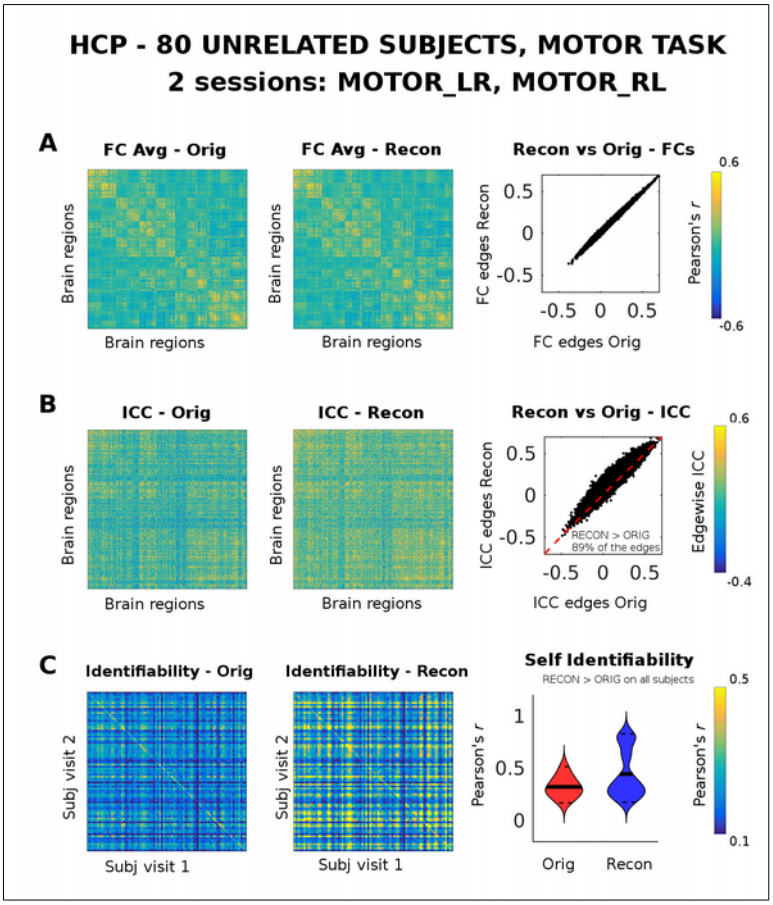


Figure 4. Evaluation of optimal PCA-reconstruction (m* = 120) on motor task-based functional connectomes (FCs) . The motor task-based FCs of 80 subjects were reconstructed at the optimal number of PCA components (i.e. 120, see Fig.2), for 1 single run. The PCA reconstruction was evaluated on the tfMRI_MOTOR_LR and tfMRI_MOTOR_RL sessions. **A1)** From left to right: the group averaged FC of the original data; the group averaged FC of the reconstructed data; the scatter plot edge by edge of the reconstructed group averaged FC (y axis) vs original group averaged FC (x axis). **B1)** From left to right: the intra-class correlation (ICC), computed over each FC edge, for the original data; the edgewise ICC for the reconstructed data; the scatter plot edge by edge of the reconstructed ICC values (y axis) vs original ICC values(x axis). The inset reports the percentage of edges where ICC increased after reconstruction (black dots on top of the red dashed line) from those that did not (black dots below of the red line). **C1)** From left to right: Identifiability matrix (i.e. the Pearson's correlation between functional profiles between subjects and sessions, see Methods) of the original data; identifiability matrix of the reconstructed data; violin plot of the "self identifiability"(i.e. the main diagonal of the identifiability matrix) distribution across the 80 subjects, for original (Orig, red) and reconstructed (Recon, blue). The solid black lines depict the mean value of the distribution; the dashed black lines the 5 and 95 percentiles. Note that self identifiability of each subject always improves after PCA reconstruction.

Self identifiability in tasks correlates with motion displacement

Further inspection of self identifiability of subjects reveals differences, mainly between resting-state and the fMRI tasks. The violin plot distributions of self identifiability values after reconstruction for resting are unimodal, whereas task sessions display bimodal and

sometimes even trimodal shapes (Fig. 3C, supplementary Fig. 2). This suggests that subject's identifiability might relate to the goodness of his test-retest FCs. That is, if one or more of the individual sessions is highly compromised or corrupted (due to motion or other factors), it might become very challenging to improve the similarity of the FCs of the same subject. Indeed, self identifiability values after reconstruction for the task sessions negatively correlate with the mean absolute motion displacement from the first fMRI frame (AbsoluteRMS, see Methods), particularly with the maximum AbsoluteRMS between the two sessions (supplementary Fig. 4). That is, the higher the presence of motion in at least one the sessions of the subject, the lower the subject identifiability. Notably, the log-linear trend between identifiability and AbsoluteRMS is evident across all the different tasks, but it is not present for the resting-state session (supplementary Fig. 4). Moreover, no significant correlation between self identifiability and task response time nor between identifiability and task accuracy (see Methods) was observed for any task (results not shown).

Reconstructed Functional connectomes: edgewise subject/task identifiability

Finally, we tested the edgewise identifiability properties of the optimally reconstructed functional connectomes, on all subjects and tasks, in two different ways. One, by computing the edgewise subject identifiability over all tasks (i.e., intraclass correlation across subjects, see Methods). In other words, identifying which specific set of functional connections can separate between subjects regardless the tasks (resting state included). The other, by computing the edgewise task identifiability over all subjects (i.e. intraclass correlation across tasks, see Methods); that is, which specific functional connections can separate between tasks (resting state included) regardless which subjects are performing them (see Methods). Interestingly, there are sub-networks that are associated to subject and task identifiability more than others, both in terms of within- and between- RSNs (Fig. 5). Particularly, edges involving connectivity within the visual and dorsal attentional networks, as well as edges between those two networks, appear among the top 5 sub-networks highly implicated in both subject and task identifiability (Fig. 5B,C,E,F). The Fronto-parietal network and its between-network connectivity with the dorsal attentional network is highly involved in subject identifiability; on the other hand, edges in somatomotor and ventral attentional have high capacity of differentiating between tasks (Fig. 5B, E).

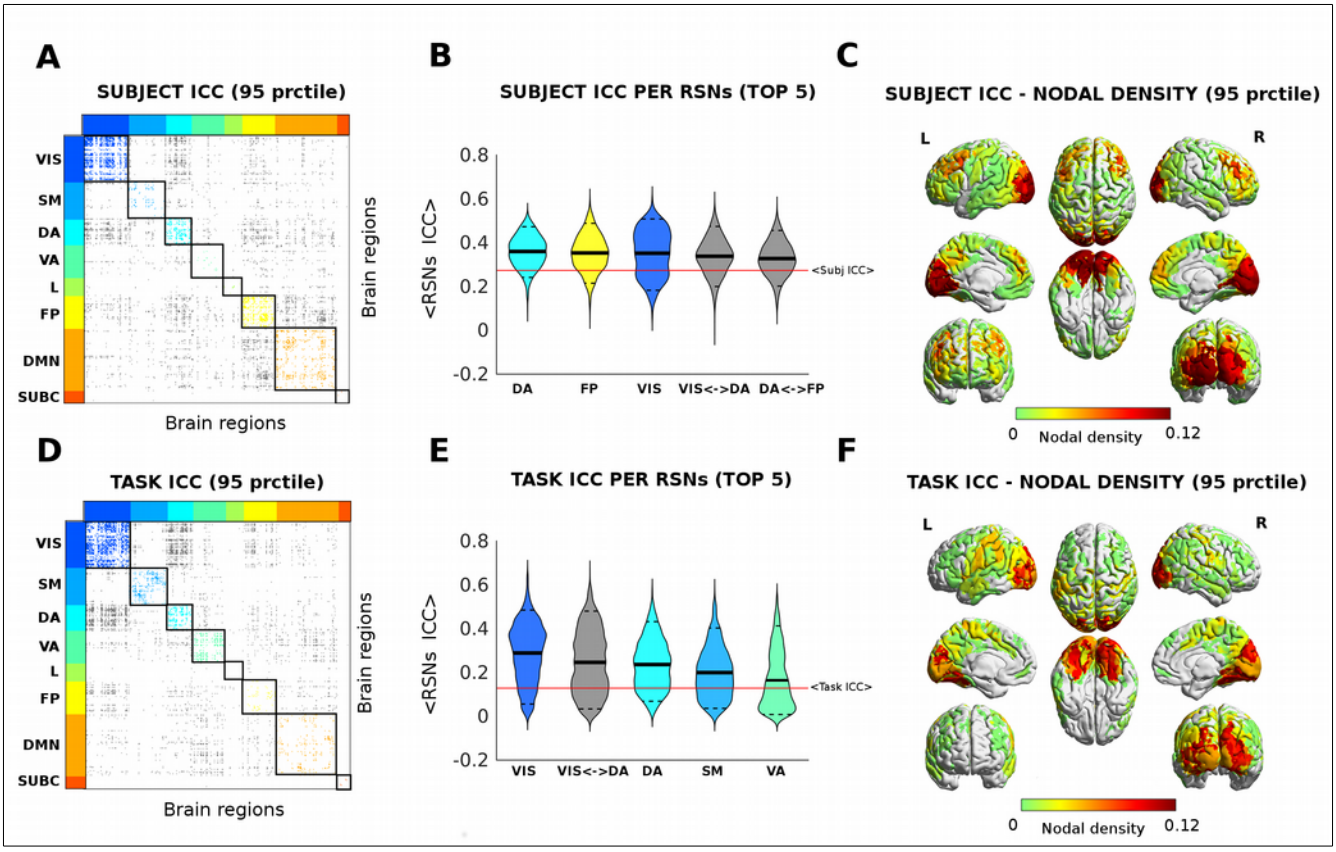


Figure 5. Intra-class correlation (ICC) analysis of subject identifiability and task identifiability. A-D) Edgewise ICC for identifiability. Figure shows functional connections for which ICC values were significantly higher (> 95th percentile of the distribution). The brain regions are ordered according to Yeo's (Yeo et al., 2011) functional resting state networks (RSNs): Visual (VIS), Somato-Motor (SM), Dorsal Attention (DA), Ventral Attention (VA), Limbic system (L), Fronto-Parietal (FP), Default Mode Network (DMN), and subcortical regions (SUBC). The colored dots depict ICC value across subjects in different within RSNs networks; gray dots indicate significant ICC edges between RSNs. The most prominent networks for subject's identifiability (5A) appear to be: VIS, and the interaction VIS-DA; FP, DMN and the interaction FP-DMN. For task identifiability (5B): VIS, and the interaction VIS-DA; DA, SM and the VA. **B-E) Violin plot of edgewise ICC for the top 5 RSNs.** The edgewise ICC distribution per within or between RSNs interaction, for the five with the highest ICC value for identifiability. Each different color indicates a different within RSN (as in 5A-D), while the gray indicates ICC values between RSN networks. The solid black lines depict the mean value of the distribution; the dashed black lines the 5 and 95 percentiles; the solid red line indicates the whole-brain mean ICC value. **C-F) Brain render of ICC subject identifiability as nodal density per region.** The strength per brain region computed as sum of edges above the 95 percentile threshold divided by the total number of edges per region gives an assessment of the overall prominence of each brain region for subject's and task identifiability. Note how, the occipital lobe is prominent in both task and subject identifiability, while frontal areas show higher nodal density for subject's ICC, as opposed to dorsal areas in task ICC.

Discussion

The neuroscientific community is advancing towards the era of large public data repositories (such as the Human Connectome Project [30], or the 1000 Functional Connectomes Project [31], the era of reproducibility of brain data and neuroscientific results [32], and the exciting avenue of linking large-scale brain connectivity profiles to single subject's genetic [33], clinical,

demographical and behavioral features [5,34]. In this respect, improving the reliability and robustness of individual fingerprinting in the connectivity domain (both functional and structural) is a crucial next step in the area of brain connectomics.

We here presented a framework that addresses this point from a cohort-level perspective. We used principal component analysis to decompose and optimally reconstruct functional connectomes obtained from the 100 unrelated subjects cohort (two sessions, test-retest) from the HCP benchmark, both in resting state and all seven fMRI-tasks. Results indicate that this method improves, on a data-driven fashion, both the global and the local (edgewise) individual fingerprint (as measured by identifiability) of the functional connectivity profiles, independently from the acquired task.

PCA is a method commonly used to provide a simpler representation of the data at hand, by compressing most of the variance of the data in a reduced number of orthogonal components, or eigenvectors. For instance, in face recognition problems, the retained eigenvectors ("eigenfaces") are used to denoise the initial images and improve the identification of the face in the image [35]. Similarly, we here mapped and ranked the principal connectivity modes from a set of functional connectomes. Hence, by maximizing individual identifiability, the reconstructed functional connectomes provided a denoised or more accurate version of the original ones.

The simplicity of the approach allowed us to test the optimal number of eigenmodes or components to retain for an optimal individual identifiability of the functional connectomes (Fig. 2) across different tasks. Interestingly, the optimal number of connectivity modes employed for the reconstruction varies across tasks (Fig. 2) whereas the variance kept was very stable. This might be due to the different complexity in the FCs inherent to the different task, which is possibly related to how engaging a task is for a human brain. Indeed, the resting state is the fMRI session where the least number of PCs is needed. Also, when splitting the resting-state time-series into two halves, each resting state run shows a lower optimal m* (m*=80 PCs for split data, supplementary Fig. 3; as opposed to m*=90 for full data, Fig. 2). We argue that this might be due to having reduced the dimensionality of the dataset when splitting of time series. Also, resting-state is the session which shows greater improvement after optimal PCA reconstruction (Fig. 3) in terms of edgewise ICC and self identifiability. However, even though the tasks include more connectivity modes to be optimally reconstructed (which might suggest a higher cognitive demand and hence farther recruiting of functional sub-circuits), the improvement in the individual identification was also substantial for all of them (Fig. 4 and supplementary Fig. 2).

We found motion (as measured by absolute frame displacement, see Methods) to be significantly associated with individual identifiability (negatively correlated) for all tasks and, interestingly, not for resting-state (supplementary Fig. 4). As mentioned above, resting-state had the lowest number of components used for the optimal reconstruction, which also means that had the highest number of components left out. This might lead to two main considerations. One, that in resting state the principal components discarded may be successfully carrying most of the motion-induced artifacts left over in the functional connectome domain. The second, that motion during resting-state sessions might be more homogeneous between subjects or better isolated in the discarded components (ranked by explained variance, see discussion on limitations below) than when the subject is cognitively engaged in a task that include time-controlled events and interactions. Ultimately, the more brain connectivity modes retained, the more fine-grained information (i.e. variance explained) is kept in the functional connectomes after optimal reconstruction. However, this comes at the

risk of carrying over session-specific and/or motion-induced connectivity artifacts, not beneficial for individual identifiability.

We extended the question to whether it is possible to identify a subject or a task performed based solely on characteristic functional connectivity patterns. Our results based on intraclass correlation show some RSNs specialization, being some RSNs more involved than others in both edgewise task-identifiability and subject-identifiability (Fig. 5). This is in line with recent studies reporting that individual differences in many tasks can be stable trait markers [36], as well as that individual fingerprint is not homogeneous across RSNs [6]. We extend these questions on individual fingerprinting by showing here that identifiability across individuals and across tasks might overlap. Indeed, we found prominent RSNs-based connections sensitive to both subjects and tasks being performed (Fig. 5).

This study has some limitations. The optimal number of components is dataset dependent (e.g., size of the cohort, heterogeneity within the cohort, acquisition and processing characteristics) and cannot be easily extrapolated from one dataset to another one. The reconstruction procedure does not ensure that the reconstructed FCs will be positive semidefinite (i.e. no negative eigenvalues). Also, here we are selecting the principal brain modes based on the ranking by variance explained (as it is the normal procedure in PCA). Different selection of the optimal subset of PCs should be explored (i.e by using simulated annealing [37]).

This work adds up to the emerging new field of features extraction in brain connectomics (independent subsystem detection through independent component analysis[21,38] dimensionality reduction and connectome denoising through PCA here), that can contribute to the association of neuroimaging data with clinical/genetic biomarkers, as well as to the exploration of the underlying and latent structures and factors present in the connectome architecture of the human brain. Future studies should explore more advanced models of features extraction as well as the connections of these denoised connectivity profiles with behavior/performance/cognition. For instance, on cross-sectional studies one could try to find optimal number of components that allows for mapping demographics or cognitive performance variables, in the lack of test-retest and hence with the impossibility of measuring identifiability. Other interesting avenues also involve the application of this methodology to structural connectivity patterns, and the dependence of the reconstruction on aging through longitudinal analyses.

Conclusion

Individual fingerprinting within the functional connectivity domain is a critical attribute for further research on brain connectomics. Here, we used data from the Human Connectome Project to demonstrate that the optimal reconstruction of the individual FCs through connectivity eigenmodes maximizes subject identifiability across resting-state and all seven tasks evaluated. The subject identifiability of the reconstructed individual connectivity profiles increased both at the global and edgewise level, also when the reconstruction was imposed on additional sessions of the subjects. We extended this approach to also map the most task-sensitive functional connections. Results showed that is possible to maximize individual fingerprinting in the functional connectivity domain regardless of the task, a crucial next step in the area of brain connectivity towards individualized connectomics.

ACKNOWLEDGMENTS

Data were provided [in part] 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. This work was partially supported by NIH R01EB022574 and by NIH R01MH108467. We would like to thank Dr. Olaf Sporns and Dr. Alex Fornito for insightful discussions.

AUTHOR CONTRIBUTIONS

E.A and J.G conceptualized the study, processed the MRI data, designed the framework and performed the analyses, interpreted the results and wrote the manuscript.

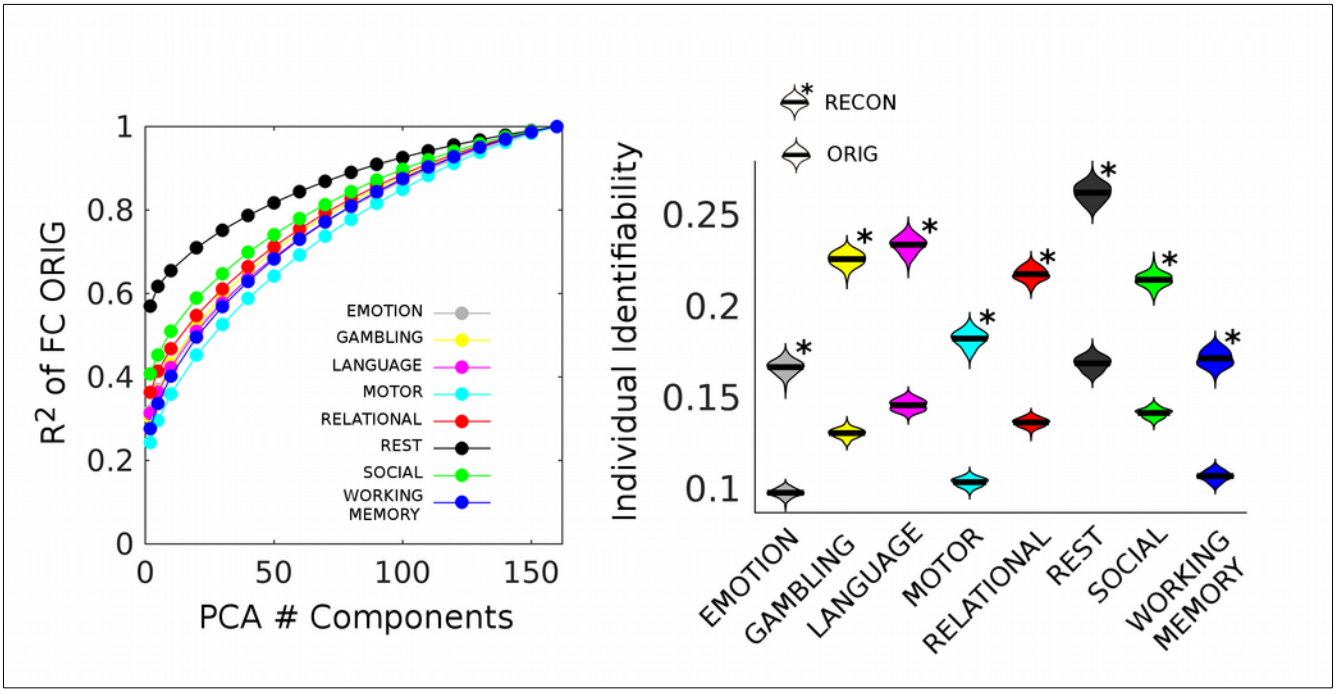
COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

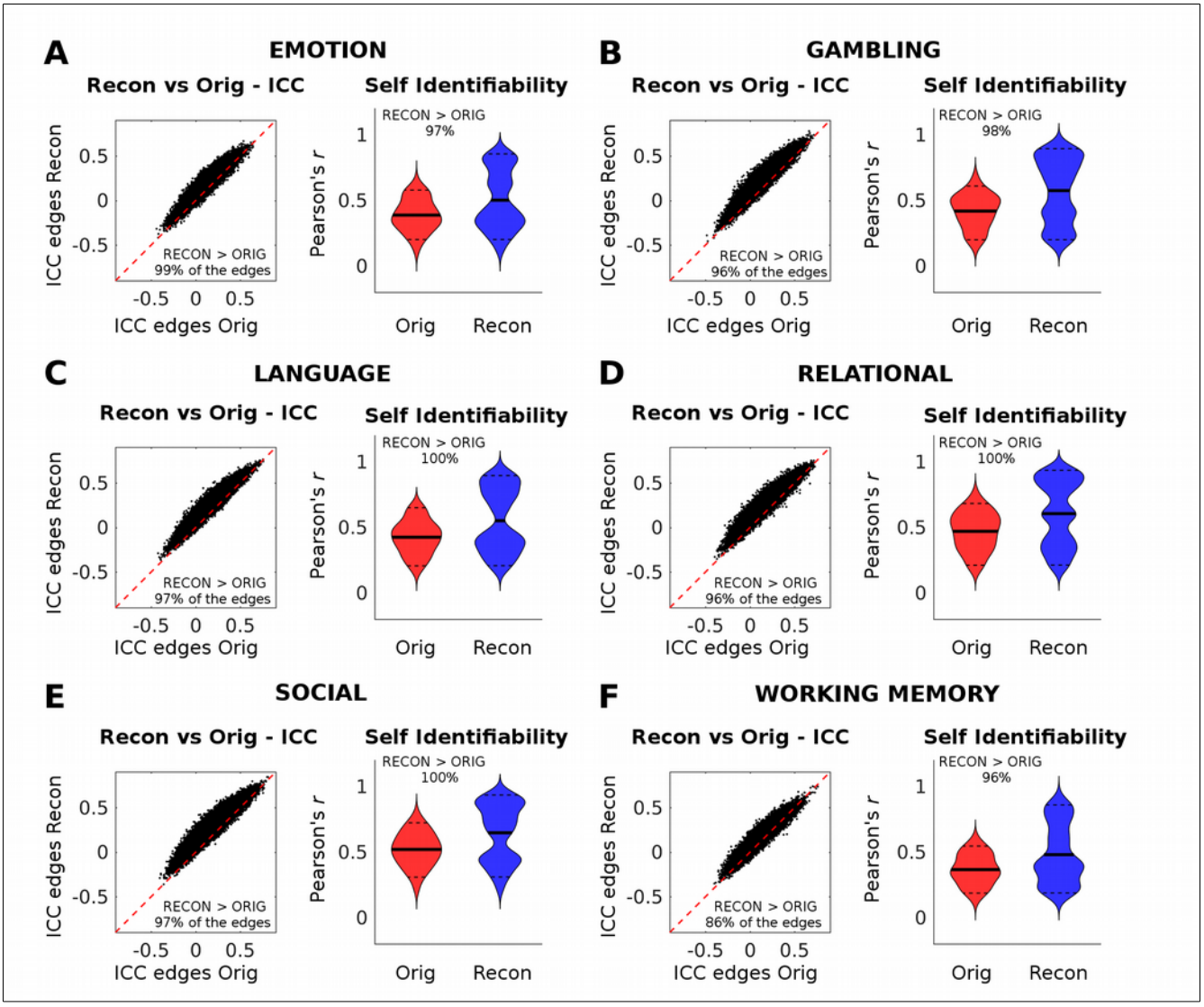
1. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10,** 186–198 (2009).
2. Sporns, O. The human connectome: a complex network. *Ann. N. Y. Acad. Sci.* **1224,** 109–125 (2011).
3. Fornito, A., Zalesky, A. & Bullmore, E. *Fundamentals of Brain Network Analysis*. (Academic Press, 2016).
4. van den Heuvel, M. P. & Hulshoff Pol, H. E. Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20,** 519–534 (2010).
5. Fornito, A., Zalesky, A. & Breakspear, M. The connectomics of brain disorders. *Nat. Rev. Neurosci.* **16,** 159–172 (2015).
6. Finn, E. S. *et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18,** 1664–1671 (2015).
7. Fox, M. D. & Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8,** 700–711 (2007).
8. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* **59,** 2142–2154 (2012).
9. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84,** 320–341 (2014).
10. Friston, K. J. Functional and Effective Connectivity: A Review. *Brain Connect.* **1,** 13–36 (2011).
11. Van Essen, D. C. *et al.* The Human Connectome Project: a data acquisition perspective. *NeuroImage* **62,** 2222–2231 (2012).
12. Smith, S. M. *et al.* Resting-state fMRI in the Human Connectome Project. *NeuroImage* **80,** 144–168 (2013).
13. Barch, D. M. *et al.* Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage* **80,** 169–189 (2013).
14. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80,** 105–124 (2013).
15. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536,** 171–178 (2016).
16. Marcus, D. *et al.* Informatics and Data Mining Tools and Strategies for the Human Connectome Project. *Front. Neuroinformatics* **5,** (2011).
17. Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62,** 782–790 (2012).
18. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. *NeuroImage* **90,** 449–468 (2014).
19. Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S. & Petersen, S. E. Intrinsic and Task-Evoked Network Architectures of the Human Brain. *Neuron* **83,** 238–251 (2014).

20. Yeo, B. T. T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106,** 1125–1165 (2011).
21. Amico, E. *et al.* Mapping the functional connectome traits of levels of consciousness. *NeuroImage* **148,** 201–211 (2017).
22. Jolliffe, I. Principal Component Analysis. in *Wiley StatsRef: Statistics Reference Online* (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat06472
23. Jackson, J. E. *A User's Guide to Principal Components*. (John Wiley & Sons, 2005).
24. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
25. Zhang, L., Dong, W., Zhang, D. & Shi, G. Two-stage image denoising by principal component analysis with local pixel grouping. *Pattern Recognit.* **43,** 1531–1549 (2010).
26. Törnqvist, L., Vartia, P. & Vartia, Y. O. How Should Relative Changes be Measured? *Am. Stat.* **39,** 43–46 (1985).
27. McGraw, K. O. & P, S. Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1,** 30–46 (1996).
28. Bartko, J. J. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychol. Rep.* **19,** 3–11 (1966).
29. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86,** 420–428 (1979).
30. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview. *NeuroImage* **80,** 62–79 (2013).
31. Mennes, M., Biswal, B. B., Castellanos, F. X. & Milham, M. P. Making data sharing work: The FCP/INDI experience. *NeuroImage* **82,** 683–691 (2013).
32. Nichols, T. E. *et al.* Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* **20,** 299–303 (2017).
33. Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N. & Nichols, T. E. Genetics of the connectome. *NeuroImage* **80,** 475–488 (2013).
34. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12,** 506–518 (2017).
35. Zhao, W., Chellappa, R. & Krishnaswamy, A. Discriminant analysis of principal components for face recognition. in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* 336–341 (1998). doi:10.1109/AFGR.1998.670971
36. Tavor, I. *et al.* Task-free MRI predicts individual differences in brain activity during task performance. *Science* **352,** 216–220 (2016).
37. Hwang, C.-R. Simulated annealing: Theory and applications. *Acta Appl. Math.* **12,** 108–111 (1988).
38. Kessler, D., Angstadt, M. & Sripada, C. Growth Charting of Brain Connectivity Networks and the Identification of Attention Impairment in Youth. *JAMA Psychiatry* **73,** 481–489 (2016).

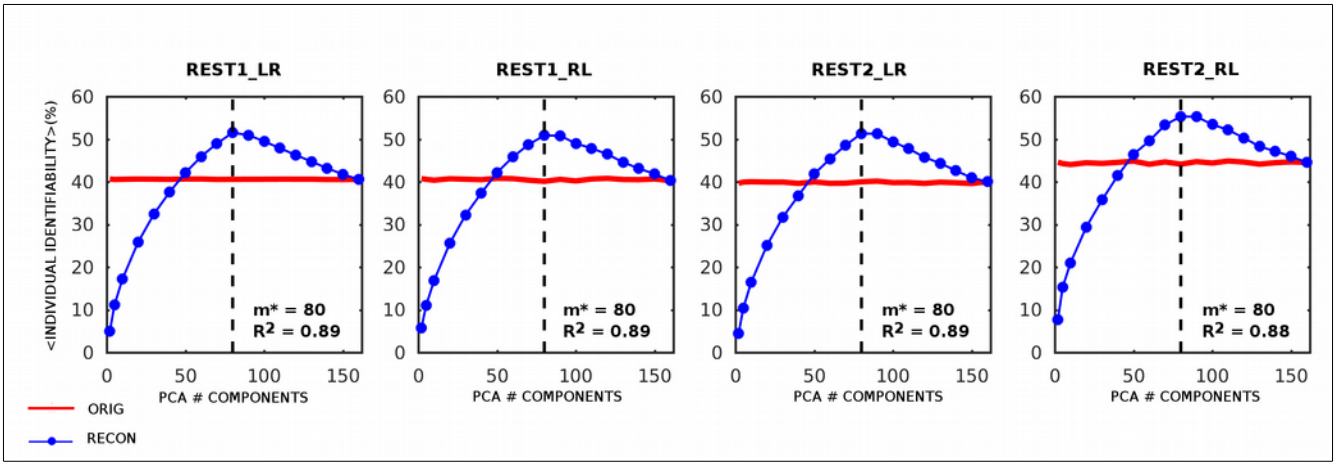
Supplementary Information



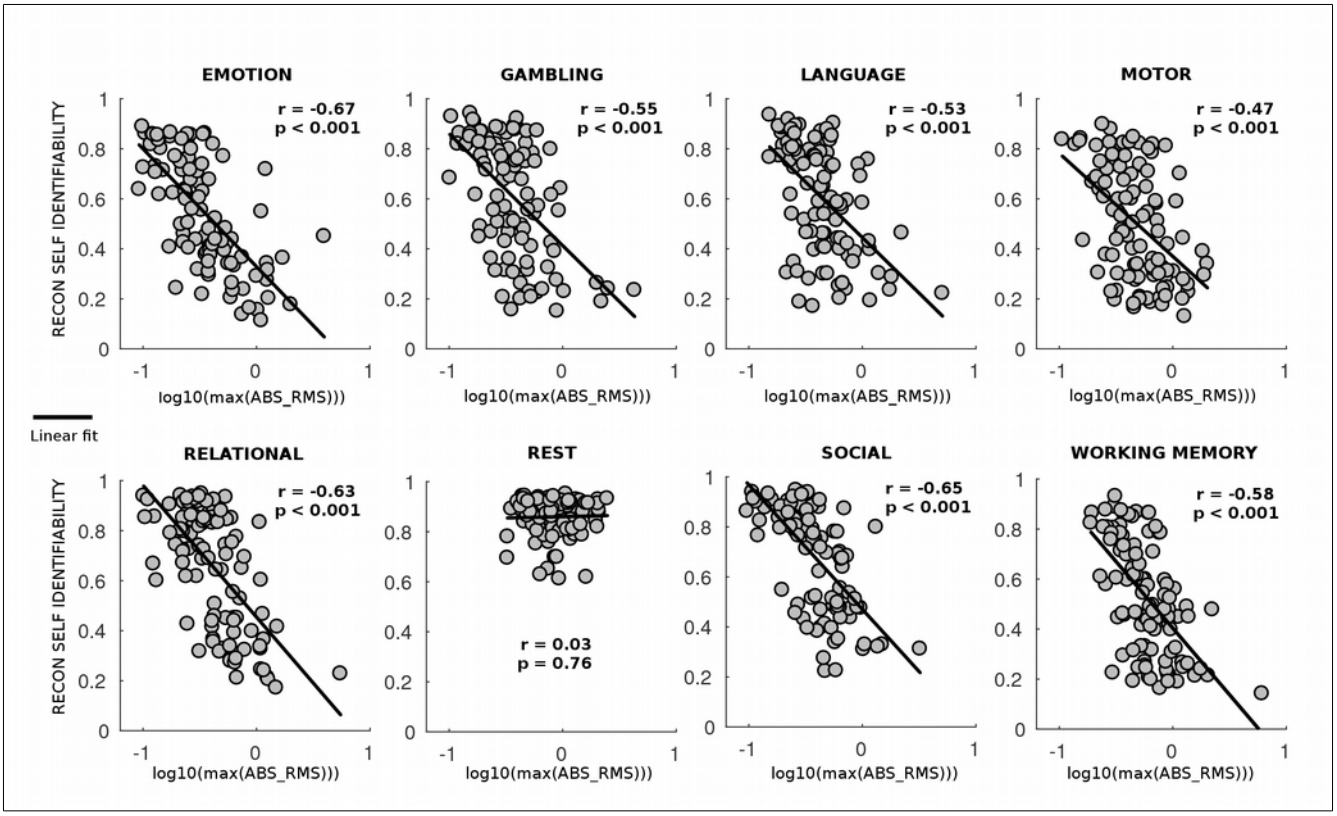
Supplementary Figure 1. Explained variance and individual identifiability across sessions. Left: the variance explained (R-squared) of the original data from the PCA reconstruction, for different number of PCA components employed. Each session is plotted with a different color. Right: violin plots show the distribution of the FC individual identifiability (see Methods) across subjects, for each fMRI session (each one has a distinct color), before and after PCA reconstruction. The solid black lines of the violins depict the mean value of the distribution. The asterisk indicates the individual identifiability distributions after reconstruction. Note how the PCA reconstruction always improves the individual identifiability.



Supplementary Figure 2. Summary of results on ICC and identifiability for the other fMRI sessions not reported in the main text. Left: the scatter plot edge by edge of the reconstructed ICC values (y axis) vs original ICC values(x axis). The inset reports the percentage of edges where ICC increased after reconstruction (top of the red dashed line) from those that did not (low of the red line). Right: violin plot of the "self identifiability"(i.e., the main diagonal of the identifiability matrix, see Methods) distribution across the 80 subjects, for original (ORIG, red) and reconstructed (RECON, blue). The solid black lines depict the mean value of the distribution; the dashed black lines the 5 and 95 percentiles. The inset specifies the percentage of subjects whose identifiability has improved after PCA reconstruction.



Supplementary Figure 3. Percent difference of the individual identifiability (I_{diff}) as a function of the number of PCA components used for reconstruction in split resting-state sessions. Plots show, for each split resting state sessions (test = first 600 fMRI frames, retest = second 600 fMRI frames, see Methods for details), the normalized individual identifiability as a function of the number of PCA components used for reconstruction (evaluated at 2, 5, and 10 to 160 components in steps of 10). Red line denotes the individual identifiability for the original FCs, whereas blue line with circles denotes the identifiability for reconstructed FCs based on the different number of components sampled. For each subplot, the optimal number of components that maximizes normalized individual identifiability (m*) and the corresponding explained variance (R2) are shown. To test the stability of the method, the individual identifiability was evaluated over 100 different runs. At each run, 80 subjects were randomly sampled from the HCP resting-state data pool of 100 unrelated subjects, 4 sessions (REST1_LR, REST1_RL, REST2_LR and REST2_RL) for a total of 160 FCs at every run. The standard deviation of I_{diff} (not shown in the plots) across runs was always lower then 0.9 %, for all the sessions considered, for both original and reconstructed data.



Supplementary Figure 4. Log-linear trend evaluation between self identifiability and mean absolute frame displacement. Plot shows, for each resting-state and task session, the scatter plot between individual self identifiability (see Methods for details) values after reconstruction (y-axis) and the log10 of the maximum value (across the two sessions) of the average absolute frame displacement (ABS_RMS, x-axis). The black lines indicate the linear fit of the scatter plots, and the insets report Pearson's correlation (r) between these two variables, with correspondent p value (p). Note how there is a significant negative correlation (p<0.001) between increases in self identifiability and ABS_RMS across all tasks. The same trend is not present in the REST acquisition.