NeuroImage 157 (2017) 250–262

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Resting-state fMRI correlations: From link-wise unreliability to whole brain stability

Mario Pannunzi[a,*], Rikkert Hindriks[a], Ruggero G. Bettinardi[a], Elisabeth Wenger[b], Nina Lisofsky[b,c], Johan Martensson[b,d], Oisin Butler[b], Elisa Filevich[e,f], Maxi Becker[b,c], Martyna Lochstet[b], Simone Kühn[b,c], Gustavo Deco[a,g]

[a] Universitat Pompeu Fabra, Theoretical and Computational Neuroscience, Center for Brain and Cognition, Roc Boronat, 138, 08018 Barcelona, Spain
[b] Max Planck Institute for Human Development, Center for Lifespan Psychology, Lentzeallee 94, 14195 Berlin, Germany
[c] University Clinic Hamburg-Eppendorf, Clinic and Policlinic for Psychiatry and Psychotherapy, Martinistraße 52, 20246 Hamburg, Germany
[d] Department of psychology, Lund University, Box 117, 221 00 Lund, Sweden
[e] Department of Psychology, Humboldt Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
[f] Bernstein Center for Computational Neuroscience Berlin, Philippstr. 13 Haus 6, 10115 Berlin, Germany
[g] Institució Catalana de Recerca i Estudis Avançats (ICREA), Universitat Pompeu Fabra, Theoretical and Computational Neuroscience, Center for Brain and Cognition, Roc Boronat, 138, 08018 Barcelona, Spain

ABSTRACT

The functional architecture of spontaneous BOLD fluctuations has been characterized in detail by numerous studies, demonstrating its potential relevance as a biomarker. However, the systematic investigation of its consistency is still in its infancy. Here, we analyze within- and between-subject variability and test-retest reliability of resting-state functional connectivity (FC) in a unique data set comprising multiple fMRI scans (42) from 5 subjects, and 50 single scans from 50 subjects. We adopt a statistical framework that enables us to identify different sources of variability in FC. We show that the low reliability of single links can be significantly improved by using multiple scans per subject. Moreover, in contrast to earlier studies, we show that spatial heterogeneity in FC reliability is not significant. Finally, we demonstrate that despite the low reliability of individual links, the information carried by the whole-brain FC matrix is robust and can be used as a functional fingerprint to identify individual subjects from the population.

Introduction

Neuroimaging techniques allow for the non-invasive investigation of two main principles of brain functioning: segregation and integration. Relationships between segregate regions can be described at different spatial and temporal scales and with different imaging techniques, and they help us to understand their integrative roles. While some techniques help to describe the physical wiring between the brain regions (e.g., diffusion tensor imaging, tractography, etc), others quantify the functional relationship between activity in different regions (Friston, 2011). To date, one of the most widely adopted techniques used to characterize the functional organization of the resting brain has been functional magnetic resonance imaging. Resting-state is commonly defined as a condition in which the participant is not performing any overt task, but lies still in the scanner (with eyes closed or fixating on a cross on a screen) while not focusing on any particular thought or sensation (see e.g., Biswal et al. (1995) or the more recent report by Zuo and Xing (2014)). Functional MRI is based on the quantification of local changes in blood oxygenation through the use of the so-called blood-oxygen level-dependent (BOLD) signals (Ogawa et al., 1990), that have been demonstrated to partially reflect underlying neural activations (Logothetis et al., 2001; Logothetis, 2008; Magri et al., 2012). Functional connectivity (FC) between different brain regions is then quantified using measures of statistical dependency (see e.g. Friston (2011)), most notably the Pearson correlation coefficient.

Resting-state functional connectivity has been used to differentiate between subjects (Finn et al., 2015) and groups, drawn either from healthy or pathological populations (see for example Rosazza and Minati (2011) for a review and references therein), or between different brain states (see for example the case of learning in Guerra-Carrillo et al. (2014) and references therein). The advantages of this technique are its high spatial resolution and large coverage (Logothetis, 2008). In fact, a resting-state fMRI scan of about 5 min allows for a character-

* Corresponding author.
E-mail address: mario.pannunzi@gmail.com (M. Pannunzi).

http://dx.doi.org/10.1016/j.neuroimage.2017.06.006
Received 9 December 2016; Received in revised form 30 May 2017; Accepted 1 June 2017

ization of the functional relationships through the brain. These advantages make this technique potentially very powerful, even considering that it measures neural activity only indirectly through BOLD signal (Logothetis, 2008). The unrestricted nature of the resting-state could in fact mirror a wide range of cognitive states and operations (Christoff et al., 2009; Richiardi et al., 2011; Hurlburt et al., 2015).

Interestingly, functional connectome studies show a differential pattern of findings: on the one hand they show a very stable architecture of correlated spontaneous activity, on the other hand they indicate a high variability in the functional structure, with temporal dynamics, ranging from less than one second (Mitra et al., 2015), to days (Anderson et al., 2011; Laumann et al., 2015). A crucial factor influencing the stability of the resting-state FC is scan duration. The most common acquisition time is 5–10 min, even though recent evidence indicates the importance of using much longer scans to obtain reliable FC estimates (Anderson et al., 2011; Birn et al., 2013; Hacker et al., 2013; Laumann et al., 2015). A question that has both theoretical and practical relevance is how much data one needs to accurately and reliable estimate the FC of an individual subject (Birn et al., 2013; Laumann et al., 2015; Finn et al., 2015).

The development of biomarkers derived from resting-state BOLD-fMRI scans that are able to characterize the functional architecture of individual brains is important for cognitive as well as for clinical neuroscience. For a biomarker to be successful, it has to be reliable; as such, two conditions must be met: on one hand, it should be stable for the same subject (or condition) across different sessions. On the other hand it should substantially vary over different subjects (or conditions). The second requirement ensures that the biomarker is selective for the variable of interest, and could thus be used to effectively discern between different subjects or conditions. The principle behind the two above mentioned criteria suggests a rather straightforward way to quantify the reliability of a potential biomarker, namely by comparing the within-subject (-condition) variability with the between-subject (-condition) variability. An index commonly adopted to measure this ratio is the intra-class correlation coefficient (ICC) a measure widely used in the psychological sciences to assess test-retest reliability (Shehzad et al., 2009; Zuo and Xing, 2014). We will use the ICC as our main tool in assessing the reliability of resting-state FC.

Although numerous studies have been devoted to characterize the functional architecture of spontaneous BOLD-fMRI fluctuations, the test-retest reliability of functional indices has begun to be addressed only recently (Anderson et al., 2011; Birn et al., 2013; Hacker et al., 2013; Zuo and Xing, 2014). From the results reported in the literature, one of the main findings is that test-retest reliability of functional indices between regions of interest (ROIs), as quantified by the intra-class correlation (ICC), seems to strongly vary over brain regions and over pairs of brain regions (for link-based indices). What has not been made explicit in previous studies, however, is an analysis of the variability of the reliability measures themselves. Indeed, reported variation of reliability has been interpreted to reflect differences in the reliability of the functional indices, without taking into consideration the statistical uncertainty due to finite sample in the estimates of the ICC.

Within the context of resting-state BOLD-fMRI experiments, in which the number of subjects and the number of scans by subject are usually limited, the variance of ICC estimators can be considerable. Assessment of the variance of ICC estimators is particularly relevant for investigating its heterogeneity over regions, links, and networks as done in Zuo and Xing (2014). In fact, a proper assessment of the ICC variability was lacking in the above-cited studies, and as such, its claimed heterogeneity has still to be demonstrated.

In the present study, we replicated most of the analyses presented in the pioneering studies Shehzad et al. (2009); Birn et al. (2013); Zuo and Xing (2014); Laumann et al. (2015). In particular, we investigate test-retest reliability of resting-state FC, and its variability. To this aim, we use fMRI to measure the resting-state activity in a group of 6 participants, each scanned 50 times, which allows to assess intersession reliability.

The paper is divided into three main sections. In the first section, we briefly present the data. In the second section, we characterize FC variability and reliability at the link-level and in particular, we analyze how these depend on the number of samples. We repeat our analysis using two different parcellations, as it has been shown before that different parcellation can influence FC estimates (e.g., Fornito et al., 2010): one based on anatomy, (AAL, Tzourio-Mazoyer et al. (2002)), and one based on functional data (Shen et al., 2013). We focus on characterizing and quantifying the nature of FC variability by decomposing it into the variability due to finite-sample statistical fluctuations and variability that genuine dynamic FC. We conclude the second section by systematically analyzing the behavior of these components as functions of scan duration and number of sessions. In the third section, we analyze the reliability of the whole FC matrix. For this purpose, we compare FC matrices obtained in different sessions both within- and between-subject.

Materials and methods

This method section is divided into seven sub-sections. The sub-section *Data acquisition and pre-processing* refers to the presentation of the data (in the Results, see *Data description*); the sub-sections *Functional connectivity analysis, Construction of surrogate data, Test-retest reliability and Sources of variability* refer to the second part of the Results (see *Data description*); the sub-sections *Definition and estimation of functional similarity and Statistical model for multivariate Gaussian biomarkers* refer to the analysis of the reliability of the whole FC.

Data acquisition and pre-processing

Fifty eight participants were recruited. Eight of the participants volunteered to be included in the longitudinal part of the study in which they were scanned 40–50 times over the course of 6 months (2 male, mean age 29, SD= 2.6, range: 24–32). Two of the participants (one male, one female) did not find the time to continue with the study and had to be excluded from further analysis (the dataset is freely available for scientific usage under request to the author S.K.; corresponding author of Filevich et al. (2017)). We had to exclude even the last male participant, who, in contrast to the instruction received, tried to apply relaxation exercise during the scan which largely influenced the measure (see Fig. S3 in the Supplementary Material). For the analysis we used 42 sessions, for homogeneity between all the subjects. The other fifty participants (all female, mean age 24, SD=3.1, range: 18–32) were part of another study that was conducted during the same period of time and underwent scanning with the same MRI sequences only once. The participants to the longitudinal study were free of psychiatric disorder and had never previously suffered from a mental disease. The other participants reported no history of psychiatric disease over a recruitment phone interview. Other medical and neurological disorders were also reasons for exclusion. No participant showed abnormalities in the MRI. The study was approved by the local ethics committee (Charité University Clinic, Berlin). After complete description of the study, we obtained informed written consent from all participants.

Images were collected on a 3 T Magnetom Trio MRI scanner system (Siemens Medical Systems, Erlangen, Germany) using a 12-channel radiofrequency head coil. Structural images were obtained using a three-dimensional T1-weighted magnetization-prepared gradient-echo sequence (MPRAGE) based on the ADNI protocol (www.adni-info.org) (repetition time (TR) = 2500 ms; echo time (TE) = 4.77 ms; TI = 1100 ms, acquisition matrix = 256 × 256×192 mm^3, flip angle = 7 deg; bandwidth=140 Hz/pixel, 1 × 1×1 mm^3 voxel size). Functional images were collected using a T2*-weighted echo planar imaging (EPI)

sequence sensitive to blood oxygen level dependent (BOLD) contrast (TR = 2000 ms, TE = 30 ms, image matrix = 64 × 64, FOV = 216 × 216×129 mm^3, flip angle = 80 deg, bandwidth=2042 Hz/pixel, voxel size 3 × 3×3 mm^3, 36 axial slices using GRAPPA acceleration factor, 5:08 min duration).

The first 10 volumes were discarded to allow the magnetization to approach a dynamic equilibrium, and for the participants to get used to the scanner noise. Part of the data pre-processing, including slice timing, head motion correction (a least squares approach and a 6-parameter spatial transformation) and spatial normalization to the Montreal Neurological Institute (MNI) template (resampling voxel size of 3 mm x 3 mm x 3 mm), were conducted using the SPM5 and Data Processing Assistant for resting-state fMRI (DPARSF, Chao-Gan and Yu-Feng (2010)). A spatial filter of 4 mm FWHM (full-width at half maximum) was used. Participants showing head motion above 3.0 mm of maximal translation (in any direction of x, y or z) and 1.0 deg of maximal rotation throughout the course of scanning would have been excluded. This was not necessary as no participant reached these criteria. We further analyzed head motion by correlating the frame-displacement measure (FD) with the estimated FC (see text). FD is reduced to a scalar value per each volume using the formula indicated in Power et al. (2012), and then averaged over volumes.

After pre-processing, linear trends were removed. Then the fMRI data were temporally band-pass filtered (0.01–0.25 Hz); but we repeated our analysis even with temporally band-pass filter (0.01–0.08 Hz), commonly adopted to reduce the very low-frequency drift and high-frequency respiratory and cardiac noise (Biswal et al., 1995; Lowe et al., 1998). The spatially normalized data were parcellated using two atlases: the automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) and a recently proposed functional atlas (Shen et al., 2013). Results for functional parcellations and for the narrow temporal filter are qualitative very similar to the ones presented in the main text, and are only reported in the Supplementary Material (see Figs. S1 and S2).

We decided to instruct participants to close their eyes during the resting state data acquisition despite the fact that resting state acquisitions with eyes open have been shown to result in slightly higher reliability of BOLD functional connectivity (Zou et al., 2015), since the resting state data acquisition, in the longitudinal study, was part of a 30 min scanning protocol that the participants completed periodically over the course of half a year. Due to this fact the likelihood of falling asleep during scanning seemed particularly high to the authors and therefore the decision was taken to record all resting states with eyes closed and ask the participants after each scan session to report whether they slept during the resting state scan or not. We tested whether being asleep or not affect the distribution, but we can exclude this possibility (see Fig. S4 in the Supplementary Material). Although recently it has been recommended to acquire 10–20 min of resting state (Birn et al., 2013; Laumann et al., 2015), we had to constrain data acquisition to 5 min per scan as the resting state sequence was only one of several sequences acquired in the longitudinal scan sessions. Moreover these 5 mins are representative of usual scanning times in many clinical studies.

Functional connectivity analysis

Spontaneous fMRI fluctuations were characterized by their population variance σ^2. For the fMRI time-series $X = (X_1,..., X_N)$ of a given ROI, σ^2 was estimated by the sample variance, which is defined as

$$\hat{\sigma}^2(X) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2,$$

where $\overline{X}$ denotes the sample mean of X. Functional connectivity was characterized by the population Pearson correlation coefficient ρ. For random variables X and Y, ρ is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}.$$

For a pair of BOLD-fMRI time-series $X = (X_1,..., X_N)$ and $Y = (Y_1, ...,Y_N)$, ρ was estimated by the sample Pearson correlation coefficient $\hat{\rho}$:

$$\hat{\rho}(X, Y) = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(X_i - \overline{X})(Y_i - \overline{Y})}{\hat{\sigma}(X)\hat{\sigma}(Y)},$$

For a given subject and link, we obtained a series of sample correlation coefficients $\hat{\rho}_1,..., \hat{\rho}_K$, where K denotes the number of scan sessions. To test for non-zero inter-scan mean and variance of the corresponding population correlation coefficients $\rho_1,...,\rho_K$ we used the sample mean and variance, respectively, of the series of sample correlation coefficients as test statistics. p-values were obtained by approximating the respective null-distributions using appropriate surrogate data (see *Construction of surrogate data*) and corrected for multiple comparisons across links using the Benjamini-Hockberg method with a false-discovery-rate (FDR) of 5%.

Construction of surrogate data

We constructed surrogate data under the null-hypotheses of zero inter-scan FC mean and variance, based on a constrained randomization procedure first proposed in Prichard and Theiler (1994). We first describe the construction for data from a single scan session and subsequently, describe how to use it to test for zero inter-scan FC mean and variance.

Let $X = (X_1,..., X_N)$ and $Y = (Y_1,...,Y_N)$ denote BOLD-fMRI time-series from two different ROI's, where N denotes the length of the scan. To construct a surrogate copy of the pair of time-series (X, Y), the discrete Fourier transforms $\widetilde{X} = (\widetilde{X}_1,..., \widetilde{X}_N)$ of X and $\widetilde{Y} = (\widetilde{Y}_1,..., \widetilde{Y}_N)$ of Y are calculated and, subsequently, the Fourier coefficients are multiplied by random (complex-valued) phases:

$$\widetilde{X}_n^{surr} = \widetilde{X}_n e^{i\phi_n^X},$$

for $n = 1,..., N$ and similarly for Y. The phases $\phi_1^X, ..., \phi_N^X$ are independently drawn from the uniform distribution on the interval $[0,2\pi]$. Surrogate copies X^{surr} and Y^{surr} of X and Y, respectively, are then obtained by applying the inverse discrete Fourier transform to $\widetilde{X}^{surr}$ and $\widetilde{Y}^{surr}$.

There are two cases to consider. In the first case, the phases ϕ_n^X are drawn independently from the phases ϕ_n^Y, and therefore the surrogate time-series X^{surr} and Y^{surr} have the same sample autocovariance functions as X and Y, respectively, but are uncorrelated. This data can hence be used to test for non-zero FC. In the second case, $\phi_n^X = \phi_n^Y$, so that X^{surr} and Y^{surr} have the same sample autocovariance functions as X and Y, respectively, but also the same sample cross-covariance function. This means that the sample correlation between X and Y is preserved and this surrogate data can hence be used to test for dynamic FC (Hindriks et al., 2015). We refer to these two types of surrogate data as *incoherent* and *coherent*, respectively.

To construct surrogate data under the null-hypothesis of zero inter-scan FC variance, we concatenated, for a given subject, the fMRI data from all scan sessions, generated 1000 coherent surrogate copies, and subsequently calculated the test-statistic values to approximate their null-distribution and to calculate p-values. Concatenating data from different sessions can lead to jumps in the time-series, and therefore to a possible bias in the statistical hypothesis testing. To exclude any bias, we assessed the performance of the testing procedure by generating 1000 synthetic data-sets with the same dimensions and a similar auto-correlation structure as the fMRI data, applied the procedure to test for non-zero inter-scan FC variance using $\alpha = 0.05$, and calculated the percentage of false positives, which yielded 5.6%. When the scan sessions were shortened, the percentage of false positives remained between 5% and 6%, only increasing to 8% in the extreme case of 15

samples per scan session. This shows that the testing procedure does not lead to excessive false positives.

Test-retest reliability

Test-retest reliability of the functional indices was quantified by the intraclass correlation coefficient (ICC), which, for a given functional index ν, in defined in terms of a random effects model (Shrout and Fleiss, 1979). Let ν_{ij} be the measured index values of subject i and scan session j, where $i = 1,\ldots,n$ and $j = 1,\ldots,k$. Thus, the index is assumed to have the following form: $\nu_{ij} = \mu + b_i + w_{ij}$, for where $i = 1,\ldots,n$ and $j = 1,\ldots,k$, and where μ denotes the expectation value of ν_{ij}, b_i denotes the random effect of the subjects, and w_{ij} denotes all residual noise (due to dynamics, measurement error or conditions/sessions). The random variables b_j and w_{ij} are assumed to be independent and normally distributed with zero mean and variance $\sigma_b{}^2$ and $\sigma_w{}^2$, respectively. The ICC of υ is now defined as

$$r = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2},$$

The ICC ranges between 0 and 1 and quantifies the test-retest reliability of the index υ. Note that for an index to be reliable, it must vary between subjects (high between-subject variance $\sigma_b{}^2$) and it must be stable across scan sessions (low within-subject variance σ_w). The most straightforward and commonly used estimator of r, which is sometimes referred to as the analytical estimator, is defined as

$$\hat{r} = \frac{BMS - WMS}{BMS + (k-1)WMS},$$

where *BMS* and *WMS* denote the mean between- and within-subjects sum of squares, respectively Atenafu et al. (2012). Although there are other estimators for r, most notably, the maximum likelihood (ML) and restricted maximum likelihood (ReML) estimators, we found them to have similar variances and only slightly different biases. The only advantage of these other estimators is the absence of negative estimates. Due to its simplicity and widespread use, we preferred to use the analytical estimator. Statistical hypothesis testing was done using an *F*-test. Specifically, since *BMS* and *WMS* are sample estimators of $k\sigma_b^2+\sigma_w^2$ and σ_w^2, respectively, the random variable

$$f = \frac{BMS}{k\sigma_b^2 + \sigma_w^2} \Big/ \frac{WMS}{\sigma_w^2},$$

is F-distributed with parameters $n - 1$ and $n(k - 1)$. Under H_0, f takes the following form: $f = \frac{BMS}{WMS}\frac{1 - r_0}{1 + (k-1)r_0}$,

which can be used to obtain the null-distribution of $\hat{r}$.

Sources of variability

The issue of finite-sample variance of the sample Pearson correlation coefficient can be assessed by the phase-randomized surrogate data (see *Construction of surrogate data*). We model the Fisher-transformed sample Pearson correlation coefficient of participant i and scan, j, denoted by $\hat{\rho}_{ij}$, with a normally distributed variable of the following form:

$$\hat{\rho}_{ij} = \bar{\rho} + \sigma_b b_i + \sigma_w w_{ij} + \sigma_f f_{ij},$$

where b, w, and f are independent, and standard-normally distributed random variables. The random variable w models the true variability in FC (within-subject), the random variable b models the between-subject FC variability, and the variable f models the finite-sample error. Assuming σ_w to be independent of i (subject) means that the true between-scan variability of FC as measured by σ_w, is the same for all subjects.

The three sources of variability can then be separated and the corresponding variances, σ_f^2 , σ_w^2, and σ_b^2, can be calculated from the surrogate analysis:

$$\sigma_b^2 = (BMS - WMS)/k,$$

$$\sigma_w^2 = WMS - WMS_0,$$

$$\sigma_f^2 = WMS_0,$$

where *WMS* and *BMS* denote the within- and between-subject mean square errors, WMS_0 denotes the mean square error within subjects for the surrogate case, and k denotes the number of sessions. These equalities allow the ICC to be written as follows:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2 + \sigma_f^2}.$$

Since the surrogate data is constructed under the null-hypothesis of zero inter-scan FC variability (that is, $\sigma_w = 0$), the ICC constructed from the surrogate data is given by

$$ICC_0 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_f^2},$$

and therefore, $ICC_0 \geq ICC$, so that the surrogate data can be used to estimate the uncertainty in the ICC that is due to the finite-sample size.

Definition and estimation of functional similarity

Central to the analysis in Finn et al. (2015) (but see also Mueller et al. (2013)) are the within- and between-subject *similarity indices*, here denoted by $\hat{R}_b$ and $\hat{R}_w$, respectively. R_w [add hat] can be calculated for every subject i and for every pair of scan sessions (j, j') and is defined as the sample Pearson correlation coefficient between the respective vectorized (and z-scored) FC matrices X_{ij} and $X_{ij'}$ with $j \neq j'$:

$$\hat{R}_w = \frac{\langle X_{ij} - \overline{\mu}_{ij} e_m,\, X_{ij'} - \overline{\mu}_{ij'} e_m \rangle}{\sqrt{\|X_{ij} - \overline{\mu}_{ij} e_m\|^2 \, \|X_{ij'} - \overline{\mu}_{ij'} e_m\|^2}},$$

where μ_{ij} is the average X_{ij} over links, $e_m \in \mathbb{R}^{m\times 1}$ denotes the vector containing all ones, and where we have suppressed the dependence of $\hat{R}_w$ on (i, j'), and j' from the notation. Similarly, $\hat{R}_b$ can be calculated for every two subjects i and i' $(i \neq i')$ and every pair of scan sessions:

$$\hat{R}_b = \frac{\langle X_{ij} - \overline{\mu}_{ij} e_m,\, X_{i'j} - \overline{\mu}_{i'j} e_m \rangle}{\sqrt{\|X_{ij} - \overline{\mu}_{ij} e_m\|^2 \, \|X_{i'j} - \overline{\mu}_{i'j} e_m\|^2}},$$

Note that R_w and R_b [add hats] can be used to assess the similarity not only for the vectorized FC matrix, but for *any* multivariate biomarker. Below, therefore, we let X_{ij} denote an arbitrary m-dimensional biomarker for subject i and scan session j.

To assess the properties of $\hat{R}_w$ and $\hat{R}_b$, we need to consider the respective population quantities, denoted by R_w and R_b, respectively. Below, we write X_{ij} for the estimated value of the biomarker and x_{ij} for the corresponding population value. The definitions of R_w and R_b are obtained by replacing the sample Pearson correlation coefficients in the equations for $\hat{R}_w$ and $\hat{R}_b$ by the population Pearson correlation coefficients and replacing X_{ij} by x_{ij}:

$$R_w = \frac{\mathbb{E}[\langle X_{ij} - \overline{\mu}_{ij} e_m,\, X_{ij'} - \overline{\mu}_{ij'} e_m \rangle]}{\sqrt{\mathbb{E}[\|X_{ij} - \overline{\mu}_{ij} e_m\|^2]\,\mathbb{E}[\|X_{ij'} - \overline{\mu}_{ij'} e_m\|^2]}},$$

for $j \neq j'$ and

$$R_b = \frac{\mathbb{E}[\langle X_{ij} - \overline{\mu}_{ij} e_m,\, X_{i'j} - \overline{\mu}_{i'j} e_m \rangle}{\sqrt{\mathbb{E}[\|X_{ij} - \overline{\mu}_{ij} e_m\|^2]\,\mathbb{E}[\|X_{i'j} - \overline{\mu}_{i'j} e_m\|^2]}},$$

for $i \neq i'$. To assess the properties (bias and uncertainty) of the estimators $\hat{R}_w$ and $\hat{R}_b$, we also need a statistical model for the population biomarker x_{ij}. This will be described in the next section.

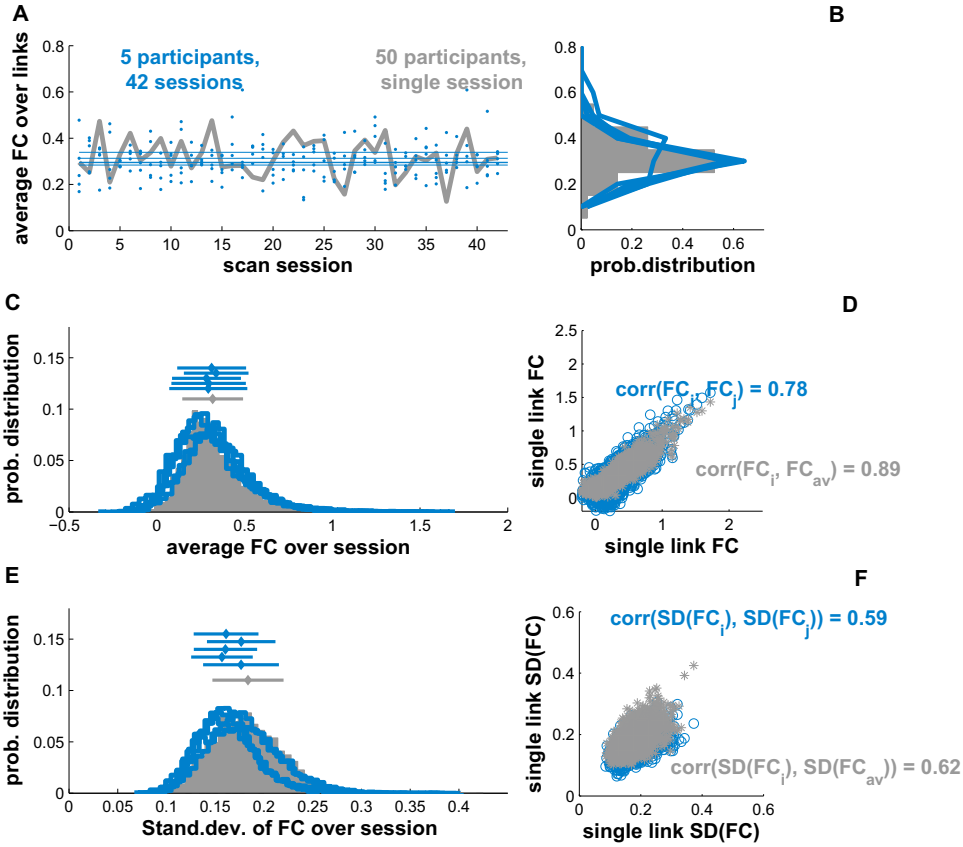

Fig. 1. Variability of FC. Panel A: Average FC over links, < FC >, of the 5 subjects for all 42 sessions (blue dots), and for the 50 subjects (gray line). Panel B: distribution of < FC > of the 5 subjects (blue lines) and the distribution of < FC > for the 50 subjects (gray bar). Panel C: the distributions of the FC values (blue lines for the five subjects, and gray bar for the 50 subjects). Panel D: the FC values of a participant (FC_i) against FC values of another participant (FC_j, blue circles), and against the FC values of the average of the 50 subjects (FC_{50s}, gray asterisks). In the same panel, we report the correlation between two participants' FC (corr(FC_i,FC_j) ≈0.8), and the correlation between a subject's FC and FC_{50s} (corr(FC_i,FC_j) ≈0.87). Panel E: distributions of the standard deviation over sessions of the FC (SD_{FC}). Same color conventions as panel C. Panel F: one participant's SD_{FC} against the SD_{FC} of another participant (blue circles), and against the SD_{FC} of the 50 subjects (gray asterisks). In this panel, we report the correlation between two participants' SD_{FC} (averaged over 42 sessions), and the SD_{FC} of one subject against the standard deviation of the 50 subjects.

Statistical model for multivariate Gaussian biomarkers

Let $x_{ij} \in \mathbb{R}^{m\times 1}$ denote an arbitrary m-dimensional (population) biomarker of subject i (i = 1,...,n) on scan session j (j = 1,...,k). In analogy to the univariate linear model used to assess link-wise test-retest reliability, we model x_{ij} by the following multivariate linear model:

$$x_{ij} = \mu + \eta_i + \xi_{ij},$$

where $\mu \in \mathbb{R}^{m\times 1}$ denotes the group-wise expectation of x_{ij}, and where $\eta_i \in \mathbb{R}^{m\times 1}$ and $\xi_{ij} \in \mathbb{R}^{m\times 1}$ denote within- and between-subject fluctuations, respectively. The random vectors η_i and ξ_{ij} are assumed to be independent and have expectation zero (that is, the m-dimensional zero-vector) and covariance matrices Σ_w and Σ_b, respectively. Note, that Σ_w and Σ_b are the generalizations to the multivariate case of the within- and between-subject variances ${\sigma_w}^2$ and ${\sigma_b}^2$, respectively.

Assuming in first approximation that Σ_b and Σ_w are diagonal matrices, the expectations of the similarity indices R_w and R_b can be expressed in terms of the model parameters as

$$\mathbb{E}[\hat{R}_w] = \frac{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b)}{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b + \Sigma_w)},$$

and

$$\mathbb{E}[\hat{R}_b] = \frac{\|\mu - \bar{\mu} em\|^2}{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b + \Sigma_w)},$$

where ***tr*** denotes matrix trace and $\bar{\mu}$ denotes the average value of μ. As a special case, suppose that Σ_b and Σ_w are identity matrices multiplied by a factor, that is $\Sigma_b = \sigma_b^2 I_m$ and $\Sigma_w = \sigma_w^2 I_m$ for certain σ_b and σ_w. Then the expressions for $\mathbb{E}[\hat{R}_w]$ and $\mathbb{E}[\hat{R}_b]$ reduce to

$$\mathbb{E}[\hat{R}_w] = \frac{\sigma_\mu^2 + \sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_w^2},$$

$$\mathbb{E}[\hat{R}_b] = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_w^2},$$

where we have defined $\sigma_\mu^2 = \|\mu - \bar{\mu} em\|/m$.

One can derive the following approximate formulas for the expectation of the similarity indices, for the more general case:

$$\mathbb{E}[\hat{R}_w] = \frac{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b) - e_m^t \Sigma_b e_m}{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b + \Sigma_w) - e_m^t (\Sigma_b + \Sigma_w) e_m},$$

and

$$\mathbb{E}[\hat{R}_b] = \frac{\|\mu - \bar{\mu} em\|^2}{\|\mu - \bar{\mu} em\|^2 + tr(\Sigma_b + \Sigma_w) - e_m^t (\Sigma_b + \Sigma_w) e_m},$$

where the variances of the similarity indices have been approximated by using Equation 3.1 in Dutilleul et al. (1993):

$$Var[R_b] = \frac{trace(B\Sigma_{wb} B\Sigma_{wb})}{trace(B\Sigma_{wb})^2},$$

and

$$Var[R_w] = \frac{trace(B\Sigma_w B\Sigma_w)}{trace(B\Sigma_w)^2},$$

where $\Sigma_{wb} = \Sigma_w + \Sigma_b$, $B = (Im - Jm/m)/m$, and Im and Jm denote m-by-m

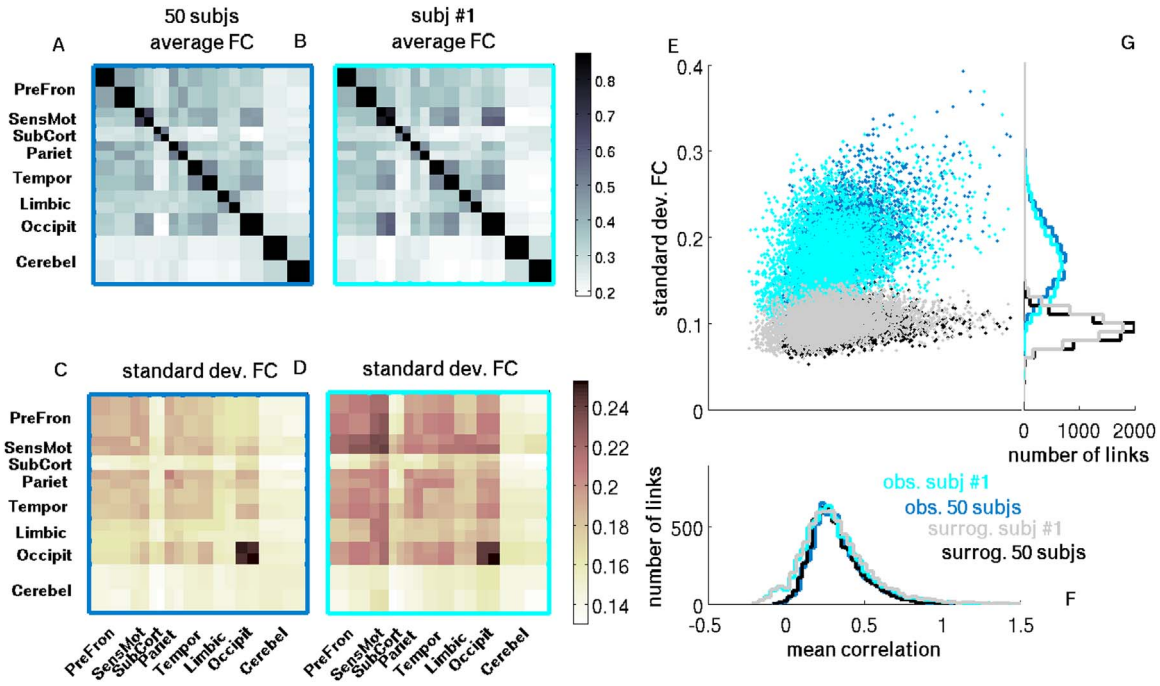

Fig. 2. Between- and within-subject FC variability. Panels A-B show the heat-maps of the average FC for 50 subjects and a single subject, respectively. Panels C-D show the FC standard deviation (SD) for 50 subjects and for a single subject, respectively. The average FC is Fisher-transformed (inverse hyperbolic tangent), and the SD is calculated from these transformed values. Color convention is cyan and blue indicate 50 subjects and single subject, respectively; black and gray dots indicate surrogate data for the 50 subjects and the single subject, respectively. Panel E shows the scatter-plot of the average FC against the FC standard deviation. Panels F and G plot the distributions for average FC and SD with the same color conventions. All the plots of this figure refer to one exemplary participant. The figures for the other four participants are qualitatively similar, but not reported.

identity matrix and the matrix of ones, respectively. We assessed the accuracy of these approximations using simulated data and found that they provide upper bounds for the respective indices. In the simulations, we generated synthetic connectivity matrices FC_{ij} (i subjects, j sessions), with a multivariate general linear model, and replacing Σ_b^{exp} and Σ_w^{exp} by their estimates. To simulate different conditions as the ones analyzed in the fMRI data, we fixed Σ_b for several simulations and used different matrices $\Sigma_w = s_w \Sigma_w^{exp}$, where s_w denotes a multiplicative factor.

Results

We present a systematic analysis of the variability and reliability of resting-state FC, both at the level of individual links and at the level of the entire brain. We used 42 scan sessions of resting-state fMRI of 5 min data from five participants (see Materials and Methods for detailed information on participants and pre-processing). ROI-level analyses were conducted using an anatomical parcellation (AAL, Tzourio-Mazoyer et al. (2002)) and the main results were replicated using a functional parcellation (Shen et al. (2013), see Supplementary Information).

Data description

Before moving into the details of the analysis, we give a descriptive overview of the data-set to provide the reader with an intuition for how variable and reliable FC is. As a first step, we consider the inter-session variability of the average link-wise FC, denoted by $< FC >$ (see panel A of Fig. 1). For the five subjects that were scanned multiple times, the average time between the first and the last session was approximately six months. Note that the average for the five subjects scanned multiple times (blue dots) resembles that computed from the 50 subjects, each of which scanned just once (gray continuous line). The same effect can be observed in panel B, in which we can compare the distribution of $< FC >$ for the five subjects (blue lines) and the distribution of $< FC >$ for the 50 subjects (gray bar).

Panel C of Fig. 1 shows that the FC distributions of the five subjects scanned multiple times (FC_i, blue lines), and of the FC of the 50 subjects (FC_{50s}, gray bar) are very similar and have similar average values. The distributions of all FC values for the five subjects and that of the 50 subjects are in general very similar, even though the latter is narrower with a standard deviation (SD) of 0.35 compared to the former, whose SD is 0.45. Another way of measuring the similarity between FC_{50s} and FC_i is through the Pearson correlation coefficient between the vectorized matrices. In our data-set, the average correlation between any couple of FC_i equals 0.8 (SD = 0.02), and the correlation between an FC_i and FC_{50s} is slightly higher; 0.87 (SD = 0.02, see the scatter-plot of panel D). Therefore, the average FC_{50} for the 50 subjects scanned just once can be considered as representative of the FC obtained from single individuals.

To complete this preliminary description, we look at the inter-session FC variability of the five subjects and subject-by-subject variability of the FC_{50} (panel E and F). We note the high similarity between the distributions of the standard-deviation over sessions of FC_i (SD_{FCi}) and of FC_{50s} (SD_{FC50s}). However, we observe rather low correlation values between the SD_{FCi} of any two of the five subjects (0.53, SD=0.02), indicating high variability between subjects of spatial distribution of FC variability.

Link-wise analysis

Within-subject variability

We now present the link-wise analysis of reliability and variability. Within-subject mean and variability of a given link's FC were quantified, respectively, by the sample mean and standard deviation of the corresponding time-series of correlation coefficients. By repeating the calculations for each link, two matrices for each subject and for the 50 subjects were obtained, corresponding to the within-subject average FC and variability. In Fig. 2 we show these matrices as heat-maps for the 50 subjects (panels A–C, blue) and for one of the subjects (panels B-D, cyan). To obtain a more robust measure, we averaged the links over

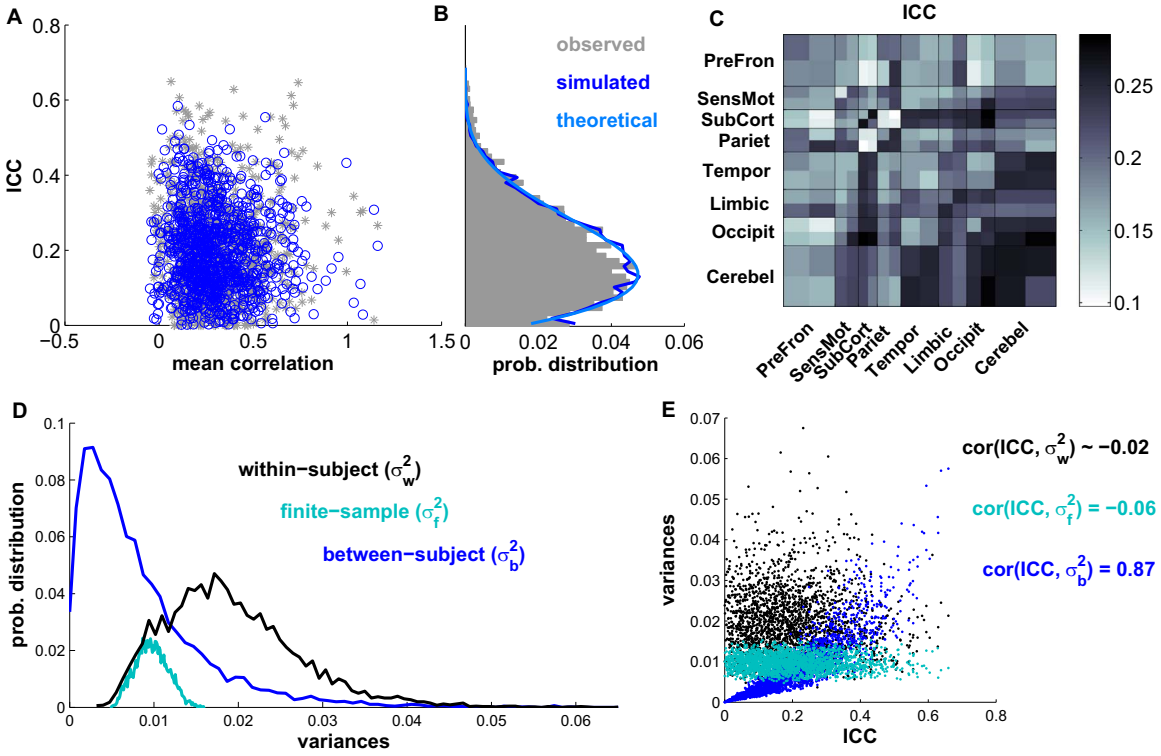

Fig. 3. Reliability of the correlation strength. Panel A shows the scatter-plot for the correlation strength against ICC value, panel B shows the histogram for the distribution of the ICC values, and panel C is the heat-map of the average ICC values for the different macro-regions. For the panels on the left and in the middle, the colors light blue, dark blue and gray refer to the theoretical, simulated and observed values, respectively (see main text). Panel D shows the distribution of the three variances (σ_f^2, σ_b^2, and σ_w^2). Panel E shows the scatter plot of the three variances against the ICC, with the values of the correlations between the three variances and the ICC; the colors follow the same convention of panel D.

macro-regions (see labels in the panels). The ROIs for this figure were defined by using the AAL parcellation (see Supplementary Material for the corresponding plots with Shen's parcellation, with very similar results). They show that both the average FC and standard deviation vary considerably over links. Note also the existence of relatively high standard deviations for some links. This suggests that the FC strength between corresponding ROI's varies considerably from scan to scan.

Panels E-G display the same average FC and its standard deviation using scatter-plot and histograms. Average and standard deviation were calculated over scans (this means over sessions for the single subject and over subjects for the 50 subjects). For both single subjects with multiple scans (cyan) and 50 subjects (blue), the standard deviation ranges between 0.1 and 0.3, with an average value of about 0.2 and a standard deviation of about 0.038 (for the single subject with multiple scans the standard deviation is slightly lower and equals 0.035). We note that the values of standard deviation of FC reported in Fig. 2E is consistent with the results reported in Laumann et al. (2015) (see Fig. S5), both for magnitude and for spatial distribution.

We can see that the value of the average FC influences the variability of the FC itself: the correlation between average FC and its standard deviation is about 0.4 for the single subject. However, this result is not robust to a global signal regression (GSR): the correlation between average FC and its standard deviation after GSR is about 0.1. Does the observed variability of the FC reflect genuine variability of spontaneous inter-areal co-activations, or does it arise from mere statistical uncertainty of the estimates? Recall that the Pearson correlation coefficients are *estimates* of the population values and as such, finite-sample variability should not be confused with the variability due to genuine underlying dynamics of the FC (see for example Lindquist et al. (2014); Hindriks et al. (2015)). With this in mind, we tested the null-hypothesis that the observed fluctuations in FC can be fully explained by statistical uncertainty of the correlation estimates: to this aim, we first constructed appropriately randomized data Prichard and Theiler (1994). This randomization method yields surrogate data with the same statistical structure and the same mean FC as the empirical data, but contains no FC dynamics (see Materials and Methods for more details). In panels E-G of Fig. 2, one realization of the surrogates is plotted (black and gray lines and circles). From panel G it is evident that the distributions of the standard deviations of the surrogate correlations (black and gray) are qualitatively different from the observed data (blue and cyan). By construction, the distributions of the mean correlation of the surrogates (blue) and of observed data (black) are identical (panel D). By repeatedly randomizing the data we can approximate the distribution of the variability for each functional connection under the null-hypothesis of constant FC variability, and hence *p*-values can be calculated. Applying the Benjamini-Hochberg method for multiple comparisons with a false-discovery rate (FDR) of 5% we found that the approximate number of functional links whose variability can be explained by the null-hypothesis of no genuine variability is around 1%. This means that practically every functional connection is dynamic for each of the five participants.

Test-retest reliability

The test-retest reliability of a measure indicates its consistency under similar conditions in contrast to dissimilar conditions: therefore, a measure is highly reliable and amenable to be a good biomarker if it yields similar results under consistent conditions, but not under dissimilar conditions. How reliable are the pairwise functional indices obtained in typical resting-state studies? To address this question, we measured the stability of the FC estimates over different scan sessions (within-subject variability) and compared them to those obtained from different participants (between-subject variability). For the functional connectivity estimates to be considered reliable, they should therefore exhibit small within-subject variability while at the same time large between-subject variability.

Following previous studies (Shehzad et al., 2009; Zuo and Xing,

2014), test-retest reliability of the functional indices was quantified by the intraclass correlation coefficient (ICC) (see *Materials and Methods*). Estimated ICC's for all links are plotted in panel A of Fig. 3 against the subject-averaged FC (gray asterisks), while panel B shows the histogram of the estimated ICC values. In the legend we indicated the gray histogram as 'observed' in contrast to the values obtained with the simulation and the theoretical analysis (see below). The heat-map in panel C shows ICC values averaged over the regions indicated in the labels.

Note that the estimated ICC values vary from link to link, ranging from approximately 0 to about 0.7. With the estimator we used (the analytic estimator), negative values of ICC can be obtained. Although maximum likelihood ICC estimates are guaranteed to be non-negative, they yielded similar results and we therefore used the analytic estimates. The average value of the ICC equals 0.22 ± 0.16, which is commonly considered rather low and indicates that link-wise FC for 5 min scan performs poorly as a biomarker for individual subjects Nunnally (1994). Despite the lack of consensus about what should be considered an acceptable level of reliability (Nunnally, 1994; Lance et al., 2006), ICC values of about 0.2 are generally considered as unacceptable. To observe such a low value, the within-subject variance has to be twice as large as the between-subject variance.

Earlier studies have reported similar values for the ICC Shehzad et al. (2009); Zuo and Xing (2014), but with substantial differences in their interpretation (see below). Similar results have also been reported by Birn et al. (2013), even though these data are more problematic as they were obtained by combining different sessions under different conditions (eyes-open, eyes-closed and fixation).

The link-wise variability of ICC estimates is also in line with previous reports that used similar scan durations Shehzad et al. (2009); Zuo and Xing (2014). In these studies, variation was interpreted as evidence for spatial heterogeneity of test-retest link-wise FC reliability (among other biomarkers), but statistical tests were not carried out. It thus remains possible that the observed ICC variability reflects statistical uncertainty, rather than true heterogeneity. Indeed, even with 42 scan sessions and 5 subjects, the variance of the ICC estimates is considerable. We therefore tested the null-hypothesis of all links having the same population ICC. The population ICC_{av} under the null-hypothesis was thus estimated by the link-wise average ICC.

We first calculated the probability of each link to have such a value of ICC or higher, given the assumption of being an estimate of ICC_{av}. This probability corresponds to a *p*-value. We subsequently calculated how many links had an ICC that was statistically different from ICC_{av} after false discovery rate correction (using Benjamini-Hochberg method with FDR = 5%).

Panel B of Fig. 3 shows the distribution of observed ICC (gray bars) and the theoretical distribution of ICC (light blue line) as estimated from a general linear model with a constant theoretical ICC (see Test-retest reliability for details). As mentioned above, the average theoretical ICC value chosen was ICC_{av}. Note that the three distributions are practically identical, which shows that there is no evidence of links that have ICC's different from ICC_{av}. The data therefore is consistent with spatially homogeneity of the link-wise FC test-retest reliability.

To further test this hypothesis, we simulated the links' correlation variability by using Gaussian variables having two sources of variability, 'within-subject' and 'between-subject'. Each simulated correlation has on average a value equal to the observed mean correlation of one real link: $FC_{sim} = FC + \sigma_w^2 \xi_w + \sigma_b^2 \xi_b$, where the variances, σ_w^2 and σ_b^2, were kept constant. The ratio between the two variances was chosen equal to the average ICC_{av}, and for simplicity we set $\sigma_w^2 = 1$ (the actual value does not influence the results of the simulation). We extracted these variables once for each simulated subject, and 50 times for each simulated scan session. Finally, we calculated the ICC values for each simulated correlation, FC_{sim}. The results of these simulations are displayed in panel A of Fig. 3 as blue circles, and their distribution in panel B as a dark blue line. Note that the simulated distributions approximate the empirical distributions quite well.

We note that this result is not sensible to GSR. After GSR the distribution of ICC is practically identical to the corresponding distribution without GSR (the average value is slightly lower).

A possible explanation for the lack of heterogeneity in the ICC values of the links might be the lack of statistical power: only few subjects, limited number of scan sessions, and correction for multiple comparisons. To circumvent the issue of having to perform a too restrictive multiple-comparison correction, we took the average ICC's of different macro-regions (the names of these regions are indicated in the labels). We use the term macro-region to indicate a brain region composed of several ROIs. The idea is based on the hypothesis that different macro-regions might have different reliabilities. This approach closely follows that taken in Zuo and Xing (2014) in which systematic differences in ICC's were reported for averaged resting-state networks, for several functional biomarkers, including (intrinsic) FC.

The heat-map in panel C of Fig. 3 shows the average correlation between pairs of macro-regions. The differences are small: the average ICCs vary between 0.1 and 0.3. We compared the ICC distribution between pairs of macro-regions with a non-parametric test (see Methods for details), and we did find most of them to be statistically different. We can therefore sort the macro-regions according to average ICC value and identify the least reliable region (parietal region, whose average ICC $\approx$ 0.15) and the most reliable region (cerebellum, whose average ICC $\approx$ 0.24). The least reliable macro-region for links connecting it to other macro-regions is the pre-frontal region (whose average ICC = 0.18) and the most reliable macro-region is the cerebellum (whose average ICC = 0.27).

Sources of variability

We now analyze the different sources of variability of the FC, and how they relate to ICC reliability. Specifically, we disentangle the contribution of three different sources of variability: 1. Genuine variability of FC in each subject (within-subject variability); 2. Variability of FC for different subjects (between-subject variability); 3. Variability of FC due to the statistical uncertainty associated with computing the correlation from a finite number of samples (finite-sample variability). The three sources of variability have already been partially accounted for in the literature (Shehzad et al., 2009; Van Dijk et al., 2010; Birn et al., 2013; Zuo and Xing, 2014; Laumann et al., 2015; Shah et al., 2016). We point out, however, that our description in terms of these variability sources is slightly different from that used in other studies (e.g., Zuo and Xing (2014); Laumann et al. (2015); Mueller et al. (2013)). As we model the correlations by random variables, each source of variability is associated with a corresponding variance: between-subject variance σ_b^2, finite-sample variance σ_f^2, and within-subject variance σ_w^2. For the sake of clarity, we note that the between-subject variance, σ_b^2, is not obtained by calculating the variances between the sessions of different subjects, that would approximated by the sum of the three variances $\sigma_b^2+\sigma_w^2+\sigma_f^2$; similarly the within-subject variance, σ_w^2, is not obtained by calculating the variances between the sessions of the same subject, that would approximately equal to the sum of $\sigma_w^2+\sigma_f^2$.

For each subject, the inter-session variability of the correlations can be divided into within-subject variability and finite-sample variability. To calculate the contribution of the finite-sample variability, we used the surrogate data described before, as they possess finite-sample variability, but (by construction) no within-subject variability (see Materials and Methods). Therefore, to obtain σ_f^2, for each link, we subtracted the value of the inter-session variability obtained from the observed data from the one obtained from the surrogate data. For the observed data, both the finite-sample variability and the between-subject variability were on average approximately half of the within-subject variability; see the complete distribution of the three variances in panel D of Fig. 3. This large difference between σ_b^2 and σ_w^2 is the main cause of the low link-wise FC reliability described in the previous section.

The values of the three variances averaged over regions are reported in Fig. S5 of the Supplementary Material. Although the variances present homogeneous values for all the regions and there is no clear pattern, we note that both the ICC and the three variances form a characteristic structure, with some macro-regions exhibiting different patterns of behavior compared to the others (see e.g., the occipital).

We also analyzed how the three variances correlate with ICC (see panel E of Fig. 3: the correlation of ICC with σ_b^2 is rather high (0.86)), while the correlations with the other two variances are almost zero, σ_w^2 (−0.07, p-value > 0.05) and σ_f^2 (−0.05, p-value < 10^{-5}). These results have a straightforward interpretation: the between-subject variability represents a structure similar to the one of the ICC, while the differences between the regions in the within-subject variability are not strongly related to regional differences in the ICC.

Recently, different studies have warned of the influence of head-motion and micro-movements (i.e. head displacements < 1 mm) on FC variability and reliability (Power et al., 2012; Laumann et al., 2016). Taking into account this possibility is indeed very relevant for our analyses, as it indicates one of the different plausible causes behind within-subject variability or (given the reliability of head motion) of the between-subjects variability. To assess this possibility, we calculated the correlation between the inter-session variability of the average frame-displacement (FD) and that of each links' correlation (see Methods for the calculus of FD). We found that head-motion explains part of the variance of the correlation (≈ 5%), even though the effect is not homogeneous (see panel A of Fig. S5 of the Supplementary Material). Moreover, the FD-effect correlates positively with within-subject variability (≈ 0.35), but not with between-subject variability.

Relevance of sample points: scan duration and multiple scans

Different studies have analyzed the effect of scan duration on FC reproducibility (Anderson et al., 2011; Birn et al., 2013; Hacker et al., 2013; Laumann et al., 2015; Finn et al., 2015), and reliability (Shehzad et al., 2009; Birn et al., 2013) demonstrating that long scan sessions increased both reliability and reproducibility. We note that the former is not an obvious consequence of the latter, in that having highly reproducible FC within-subject could also mean highly reproducible FC between-subject and therefore low reliability. For example, Birn and colleagues demonstrated that reliability slowly increased with scan duration: on average, the maximal ICC value for very long scans (30 min) is rather low, ICC ≈ 0.4 (Birn et al., 2013).

We therefore systematically studied the influence of scan duration on the reliability of FC indices. Moreover, we analyzed the behavior of the ICC as a function of the different sources of variability (within-subject, between-subject, and finite-sample). Panel A of Fig. 4 shows the behavior of the three variances and that of the ICC for different scan durations. For very short durations (below 1 min), finite-sample, σ_f^2 (green line) is the most relevant source of variability even though its contribution rapidly decreases with increasing scan duration. We observed that the behavior of the finite-sample variability can be approximated by a power law of $1/N^a$, where N is the number of time points, and a is about 1.3 (χ^2 = 0.98). This is not surprising, as the finite-sample variance of the correlation between any two time-series having zero auto-correlation equals one. Within-subject variability (blue line) also tends to decrease with increasing scan duration, even though at a much slower rate, whereas between-subject variability (black line) remains approximately constant.

We observe here (panel A), just like in panel D of Fig. 3, that the relevance of the finite-sample is decreasing toward zero. This is dependent on the number of samples (obviously), and already for 60 samples (~two minutes) the average σ_f^2 is inferior to average σ_w^2. Further studies will be needed to understand the discrepancy between these results and the results reported in Laumann et al., (2015, 2016), where it is claimed that the day-to-day variability "is almost entirely (> 98%) attributable to sampling error" (it is plausible that within "sampling error" Laumann and colleagues included both finite-sample variance and within-subject variance).

Having acquired multiple scans from the same subject enabled us to measure the three variances and the ICC obtained using the average FC over several sessions (details can be found in the Materials and Methods). Results from this analysis are depicted in panel B of Fig. 4, in order to directly compare them to the evolution of the variances for different scan durations. We note that the between-subject variability σ_b^2 again tends to remain constant, whereas the finite-sample variability σ_f^2 continues to decrease with no evident changes in slope. On the other hand, within-subject variability σ_w^2 seems to exhibit discontinuous changes that are mirrored by abrupt changes in the slope of ICC. These abrupt changes are expected, given the previous results reported in literature (Shehzad et al., 2009; Birn et al., 2013) on the difference between the reliability within-scan session (less than one hour) and between-scan sessions (more than one month), with higher values of reliability for the case within-scan session. This indicates that to obtain a higher reliability, the FC needs to be averaged over multiple sessions. Indeed, according to Birn et al. (2013), there seems to be a plateau for the ICC between-scan sessions above 18 min (see Fig. 3a of Birn et al. (2013)). Evidence for this slope change can be found in the high ICC value (0.7) obtained for FCs extracted from an average of six sessions (summing up to approximately 30 mins).

We underline the relevance of this analysis: First, we can describe how reliability of the FC changes as a function of scan duration or using

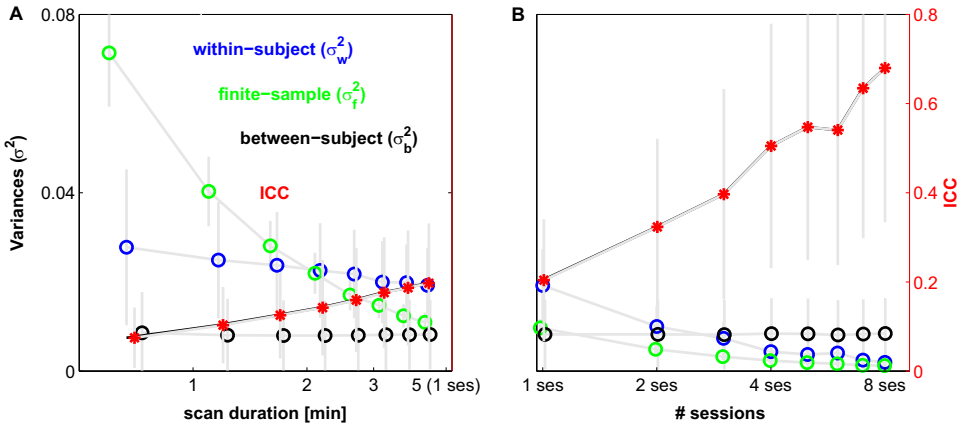

Fig. 4. Effect of scan duration on the FC reliability. The graph shows the behavior of the average reliability, ICC, and the behavior of the three variances related to the three sources of variability of FC for different scan duration (panel A) and using multiple scan sessions (panel B). The empty circles refer to the three sources of variability: within-subject (σ_w^2, blue), finite-sample (σ_f^2, green) and between-subject (σ_b^2, black). The red asterisks refer to ICC. To plot ICC we used a second y-axis (in red, on the right). In gray, the SD of each measure is reported. The points are slightly misaligned to improve the plot's readability.

several scan sessions for the three different types of variance. Second, we conclude that the use of multiple sessions seems to be a potential way to overcome the low reliability upper-limit indicated by Birn et al. (2013). The relevance of the finite-sample variability is conspicuous, but it decreases with increasing scan duration. The influence of scan duration on the sources of variability is analyzed in the next section.

Global FC analysis

After having analyzed reliability from a local, link-wise perspective (*Link-wise analysis*), we focused on studying the inter-scan variability of the whole-brain, global FC structure. This means that instead of considering the variability of the different pair-wise FC's, we considered the within- and between-subject variability of the vectorized FC matrices in their entirety. As in the local analysis (see *Link-wise analysis*), and following earlier studies Mueller et al. (2013); Laumann et al. (2015); Finn et al. (2015), functional connectivity was quantified using the Pearson correlation coefficient. The richness of information contained in the multivariate structure of whole-brain resting-state FC matrices has recently been demonstrated Finn et al. (2015). In that study, it was shown that the FC matrix can be used as a "functional fingerprint" in that it allows identification of individual subjects from a 30-min resting-state scan. The findings in Finn et al. (2015) appeared to be in stark contrast with the low test-retest reliability of local FC indices. For example, Birn et al. (2013) reported low ICC's (≤ 0.4) for pair-wise Pearson correlations even for long scan sessions (30 min) and we reported similar values (see *Link-wise analysis*). In this section we reproduce the findings in Finn et al. (2015) (*Subject identification from resting-state FC*) and provide a statistical framework that can be used to assess the factors influencing functional fingerprinting (*Quality of functional fingerprints*). Taken together, our results confirm the strength of whole-brain FC analysis over local measures.

Subject identification from resting-state FC

In this section, we reproduce the observations of Finn et al. (2015) and again assess the effect of scan duration. The analysis carried out in Finn et al. (2015) is based on the sample Pearson correlation coefficients between different pairs of vectorized FC matrices, to which they referred to as *similarity indices*. These similarity indices can be calculated between (vectorized) FC matrices of different scans of the same subject (within-subject) or between FC matrices obtained from different subjects (between-subject). The within- and between-subject similarity indices are here denoted by R_w and R_b, respectively. Details are provided in Section 2.7. Finn and colleagues demonstrated that for 30-min resting-state scans, $R_w > R_b$, for practically all values of R_w and R_b (calculated from all possible scan-pairs), which implies that R_w and R_b can be used as "functional fingerprints" to identify individual subjects. We repeated the analysis by calculating the distribution of R_w and R_b, collapsing over different sessions. In Fig. 5, panel A,C-E show the observed distributions of (Fisher transformed) R_w (gray) and R_b (black) for a different numbers of samples (number of sessions). Note that the separation between the distributions of the two similarity indices, R_w and R_b, increases rapidly with increasing number of sessions: from panel A (1 session) to panel E (6 sessions). This separation is almost complete (zero overlap between the distributions) for four sessions. This is noteworthy as with four sessions, the average ICC of single links is still around 0.5 (similar value is reported by Birn et al. (2013)). On the whole-brain level, in contrast, they allow to identify individual subjects (Finn et al., 2015).

To explain why functional fingerprinting is possible and how its quality depends on different factors, we constructed a statistical model for the vectorized (and *z*-transformed) FC matrices. Specifically, the vectorized FC matrix of subject *i* at scan *j*, denoted by x_{ij} is modeled as a normally distributed random vector having the following structure:

$$x_{ij} = \mu + \eta_i + \xi_{ij}, \tag{20}$$

where $\mu \in \mathbb{R}^{m\times 1}$ denotes the group-wise expectation of x_{ij}, and where $\eta_i \in \mathbb{R}^{m\times 1}$ and $\xi_{ij} \in \mathbb{R}^{m\times 1}$ denote within- and between-subject fluctuations, respectively. The random vectors η_i and ξ_{ij} are assumed to be independent and have expectation zero and covariance matrices Σ_w and Σ_b, respectively (see *Statistical model for multivariate Gaussian biomarkers* for more details). As we will see below, we can express all properties of the similarity indices R_w and R_b and their estimators $\hat{R}_w$ and $\hat{R}_b$ in terms of the model parameters μ, Σ_w and Σ_b.

Instead of considering R_w and R_b, it will be convenient to consider their Fisher-transformations, denoted by z_w and z_b, respectively, and similarly for their estimators. We first consider the special case in which Σ_b and Σ_w are diagonal matrices (but see panel B of Fig. 5 to see the observed values of Σ_b and Σ_w), that is $\Sigma_b = {\sigma_b}^2 I_m$ and $\Sigma_w = {\sigma_w}^2 I_m$ for certain σ_b and σ_w. In *Definition and estimation of functional similarity* it is shown that in this case, the similarity indices can be expressed in terms of the model parameters as follows:

$$R_w = \frac{\sigma_\mu^2 + \sigma_b^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_w^2}, \tag{21}$$

and

$$R_b = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_b^2 + \sigma_w^2}, \tag{22}$$

where we have defined $\sigma\mu 2 = \|\mu - \mu em\|/m$. Note that ${\sigma_\mu}^2$ is the variance of FC over links that is common to all subjects. These formulas allow interpreting the similarity indices and relating them to the link-wise ICCs, or more exactly to the parameters determining it.

Quality of functional fingerprints

Fig. 4 shows that ${\sigma_\mu}^2$ and ${\sigma_b}^2$ do not depend on the duration of the scan and that ${\sigma_f}^2$ and ${\sigma_w}^2$ decrease with increasing scanning duration. This implies that, for increasing scan duration, $z_b = \operatorname{arctanh}(R_b)$ is bounded from above while, $z_w = \operatorname{arctanh}(R_w)$ increases without bound. Furthermore, the variances of z_b and z_w are bounded from above (by one over the number of links). The consequence of these two findings (infinite separation between the average values and lower bound of the distribution variability) is that the distributions of z_b and z_w are asymptotically separated. This result follows from the fact that for every brain region i, there exists a constant FCi, while the true variability, ${\sigma_w}^2$, decreases rapidly for increasing scan duration. Thus, the global FC model qualitatively agrees with our experimental observations. It also has the advantage of being simple in that a few parameters and assumptions determine it completely. To conclude, the distributions of (Fisher transformed) R_w and R_b are Gaussian with approximately constant variances and with expectation values that diverge with a speed approximately equal to the number of samples.

To conclude, the model proposed to describe the whole FC is qualitatively in agreement with the experimental results, and it has the advantage of being simple: Few parameters and marginal assumptions determine it completely. In a nutshell: the distributions of (Fisher transformed) R_w and R_b are two Gaussian distributions, whose variances are approximately constant and whose expected values are moving away from one another tending toward infinite values, and with a speed that follows approximately the number of samples.

Discussion

In this study, we assessed the variability and test-retest reliability of the human functional connectome. To this aim we used a unique dataset comprising multiple (42) fMRI scans of five minutes each for five subjects obtained during a classical resting-state paradigm, together with another sets of single-scans obtained from 50 different subjects.

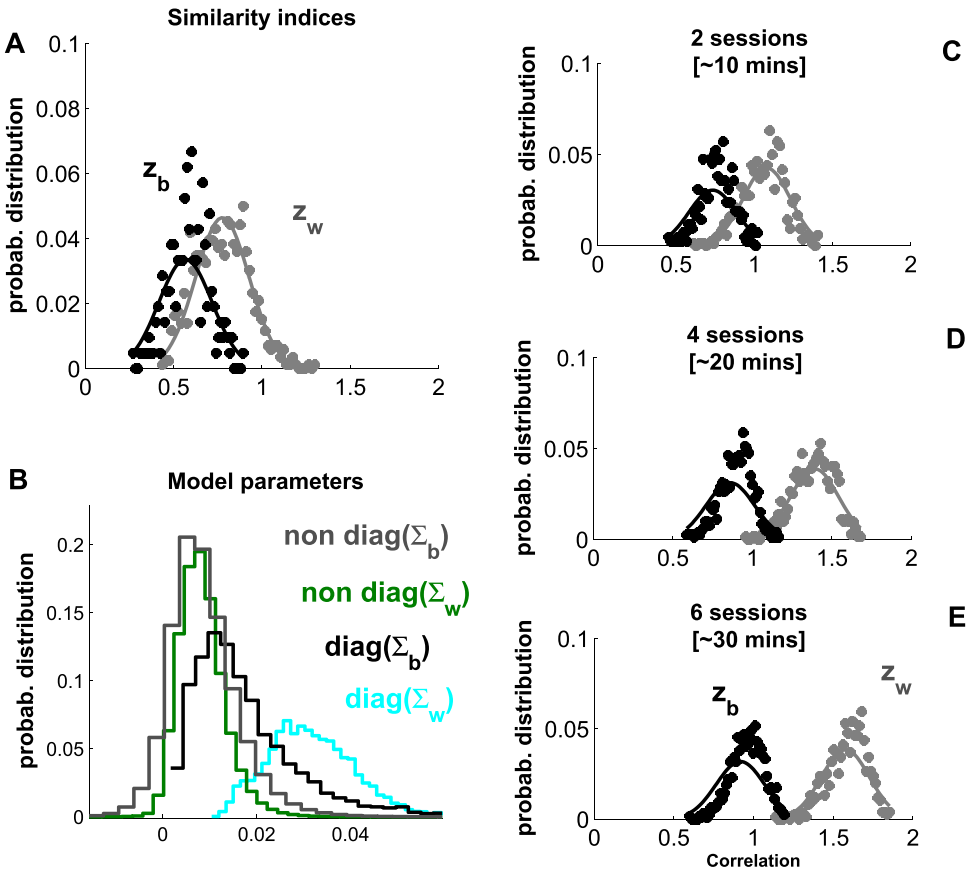

Fig. 5. Analysis of FC at global level. The two panels on the left refer to the analysis done using a single session. Panel B shows the distributions of the estimated parameters' values with the general linear model. Panels A, C-E plot the distribution of z_w (gray) and z_b (black) for FC averaged over 1, 2, 4, and 6 sessions, respectively. The observed data are represented with dots, and the theoretical approximated values with continuous lines. The separation between the distributions of z_w and z_b increases rapidly when increasing the number of sessions.

Single link reliability

We first analyzed the reliability of single functional links between ROIs, as quantified by the ICC. In order to avoid potential biases due to the parcellation, ROIs were obtained both using an anatomical (AAL) as well as a functional parcellation recently proposed by Shen et al. (2013). From our results we conclude that the average reliability of single-link FC is quite low (≈0.2) which is in agreement with the literature (Shehzad et al., 2009; Birn et al., 2013). These results, as well as all other results reported in the present work, are qualitatively equal for the two parcellations. Interestingly, we found that the correlation values of all links have an ICC drawn from the same distribution. In other words, our data support the hypothesis of an overall homogeneity in the reliability of functional links, in contrast to what has been suggested in previous literature (Shehzad et al., 2009; Zuo and Xing, 2014).

A small ICC variance is crucial to distinguish between reliable and unreliable links. To obtain a small ICC variance, a very large number of both subjects and scan sessions is needed. To date, analyses of resting-state fMRI test-retest reliability typically used a large number of subjects with two or three scans per subject. We adopted the opposite strategy, but still did not reach a lower ICC variance: we used 42 scans for five subjects, and still the standard deviation of the estimated ICC was approximately 0.2, which is similar to the values obtained from a data-set containing 75 subjects and 3 scans Zuo and Xing (2014). Therefore, to substantially decrease ICC variance in successive studies, it is necessary to use either a larger number of subjects or scans. For example, a fifth of the ICC standard deviation can be achieved with 100 participants instead of six, and 42 scans. These numbers point out the difficulty of determining FC test-retest reliability empirically. However, we recall that 42 scans and five subjects (or 75 subjects and 3 scans) are more than sufficient to obtain a good estimate of the average ICC (as it can easily be tested numerically).

Sources of variability

In this study we characterized and quantified different sources of variability in the correlation between different brain regions. Thanks to the use of surrogate data, we could effectively distinguish between three distinct sources: 1. the statistical uncertainty due to calculating correlations from a finite number of samples (finite-sample variability); 2. Genuine session-dependent variation in functional correlations between different brain regions within-subjects variability; 3. Between-subject variability.

Separating these different sources of FC variability allows us to quantify dynamic FC as well as link-to-link differences in reliability itself. For example, between-subject variability shows time-consistency, in contrast to the behavior of the finite-sample and within-subject variances, as both decrease for increasing number of sample points. While the decrease of the finite-sample variance with the number of samples is trivial, neither the between- nor the within-subject variances' behavior can be predicted from previous analyses.

Moreover, from this result, and the results of Birn et al. (2013), we can predict that within-subject variance of FC reaches a plateau, at ≈0.012. Indeed, Birn and colleagues showed that the link reliability reaches a maximum value (0.4) for scan duration of approximately 20 min. Such a low ICC value limits the use of single link FC as a potential biomarker. Here, we showed that a possible solution is to join multiple sessions to obtain an intermediate to high ICC. Here, 6–8 sessions of 5 min each are required per subject. This being said, it is

clearly more convenient to use longer scan sessions to diminish the minimal number of required scan sessions.

We showed that FC variability is in part due to the use of a finite number of samples. Thanks to our surrogate-based analysis, we could quantify the relative contributions of the finite- sample and genuine variability. We found that this genuine (within-subject) variability is, for 5 min scan session, approximately equal to the finite-sample variability. The sum of these two sources of variabilities are quantitatively in agreement with what reported in Laumann et al. (2015, 2016), and more in general our results on the ICC seem in quantitative agreement with what has been reported in other reliability studies (Shehzad et al., 2009; Birn et al., 2013).

Despite all macro-regions having approximately the same (low) ICC, there is some region-to-region variability. These differences could, in principle, be caused by any of the three variances; however, we showed that in fact it is mainly the between-subject variability that is responsible for the slight differences in ICC between different regions. For example, we found higher values of ICC in cerebellum and lower values of ICC in pre-frontal cortex. Similar analysis were carried out in Laumann et al. (2015); Mueller et al. (2013), however, in those studies the three sources of variability were not separated, which makes the results more difficult to interpret.

From link-wise unreliability to whole brain stability

We analyzed the similarity of the whole-brain spontaneous correlation structure of the same subjects across different sessions, as well as that between different subjects within a general linear model framework (for similar approaches see Mueller et al. (2013); Finn et al. (2015)). This model provides theoretical ground to understand and solve an apparent paradox: how is it possible that low link-wise reliability (Birn et al., 2013) goes together with high stability at the global level (whole FC), as has recently been shown (Finn et al., 2015)?

To this aim, we studied the distribution of two similarity indices, R_w and R_b, that measure the similarity (in terms of Pearson correlation coefficients) between FC matrices of two sessions of the same subject and of two subjects, respectively. Taking advantage of the multiple sessions of our data set, we calculated the distribution of these indices for an increasing number of concatenated sessions. In addition, we obtained an approximate expression for the average and the variance of the estimators of the distributions of R_w and R_b (see Eqs. 18 and 19). These estimators are simple functions of the between- and within-subject variances. It is straightforward to show that for an increasing number of sessions, the average Z_w *(the Fisher transform of* R_w*)* converges to infinity, while the average Z_w *(the Fisher transform of* R_b*)* remains finite and that their variances are limited. Therefore, if the two distributions do not overlap, the identification is perfect.

Finn et al. (2015) assessed the identification issue without directly adopting these similarity indices. Moreover, they used an increasing number of data-points within the same scan session instead of multiple sessions. The latter is a considerable difference, and as we mentioned before, based on the analysis of Birn and colleagues, we predict a lower asymptotic value for the within-subject variance for scan sessions longer than 20 min. This implies, following our analysis, that R_w has an asymptotic limit (upper bound) for scan duration greater than 20 min. This prediction is confirmed by the results shown in Fig. 3B in Finn et al. (2015).

The relevance of this analysis goes beyond this result: indeed, we hope that this statistical framework can be used as an general tool for analyzing FC and to connect single-link analysis with the analysis of macro-regions and the whole-brain FC.

Limitations

The resting-state literature has proposed several measures to characterize spontaneous fMRI fluctuations (see for example the review by Zuo and Xing (2014). These measures can be related to single voxels Zuo and Xing (2014)), to larger functional networks (based, for example, on independent component analysis), or to the statistical interdependencies between the time-courses of different voxels or regions.

In this study we focused on only one measure, namely the Pearson correlation coefficient obtained from the BOLD signals of different pairs of ROIs. We considered this measure as a starting point, and indeed all analyses performed here can be applied to alternative measures as well. This choice is motivated by two factors: its simplicity, a linear measure of the relationships between activities, and its widespread use in the field of resting-state fMRI. However, in the recent past, different measures of BOLD activities have been presented (see the ones analyzed in Zuo and Xing (2014)) increasing the potential of fMRI studies.

In our study, we investigated linkwise and whole-brain FC variability using two parcellations: one based on anatomy, the AAL, and one based on the functional parcellation (Shen et al., 2013). We did not find strong quantitative differences in the results of the two parcellations. However, different studies (e.g., the graph study Fornito et al. (2010)), illustrated the relevance of the parcellation. Moreover, a recent study showed possible pitfalls of these two parcellations (see e.g., Gordon et al., 2016) due to their inexact registration of functional areas. Therefore future analyses with the parcellations presented there is desirable, hopefully with a comparative analysis between the available parcellations.

We assessed FC variability without directly analyzing its origin (apart from head-motion). Other studies already started to focus on this important aspect, that can have a very broad application, going from physiological (body heat, cardiac and respiration artifacts, head motion) to technical (machine noise, scanner type, experimental instructions, data standardization, data pre-/post-processing strategies) to brain status (e.g., Rack-Gomer, Liau et al. (2009); Birn (2012); Shannon, Dosenbach et al. (2013); Yan et al. (2013); Hurlburt et al. (2015); Yan et al. (2013); Power et al. (2012); Tagliazucchi and Laufs (2014); Laumann et al. (2016)). It will be useful to capitalize on the description developed in this work, and to use these insights when planning future studies. This will likely improve our understanding of the sources of variability in the human functional connectome.

The potential of resting-state functional connectivity is well illustrated by its ability to characterize both healthy and abnormal cognitive processes and to predict perception and performance. Further drawing from its potential, however, requires a systematic assessment of its variability and test-retest reliability. Our study has demonstrated how such an assessment, together with the application of appropriate statistical concepts, helps to explain the apparent contradiction between local unreliability and global stability of resting-state fluctuations in the human brain.

The dataset is freely available for usage in scientific research. To prevent its circulation unrelated to research usage, we ask that scientists interested in obtaining the dataset email S.K., corresponding author of Filevich et al. (2017).

Acknowledgments

This research is supported by the European Research Council (ERC) Advanced Grant DYSTRUCTURE (no. 295129), by the Spanish Research Project PSI2013-42091; PSI2016-75688-P; by the Catalan Agency for Management of University and Research Grants, AGAUR (2014SGR856); (RGB) FI-DGR scholarship of the Catalan Government through the Agencia de Gestío d'Ajuts Universitari i de Recerca, agreement no. 2013FI-B1-00099.

SK is supported by a Heisenberg grant from the German Science Foundation (DFG KU 3322/1-1), the European Union (ERC-2016-StG-Self-Control-677804) and the Jacobs Foundation (JRF 2016–2018). EF is supported by a Freigeist Fellowship from The Volkgswagen Foundation (Az: 91 620).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neuroimage.2017.06.006.

References

Anderson, J.S., Ferguson, M.A., Lopez-Larson, M., Yurgelun-Todd, D., 2011. Reproducibility of single-subject functional connectivity measurements. Am. J. Neuroradiol. 32 (3), 548–555.

Atenafu, E.G., Hamid, J.S., To, T., Willan, A.R., Feldman, B.M., Beyene, J., 2012. Bias-corrected estimator for intraclass correlation coefficient in the balanced one-way random effects model. BMC Med. Res. Methodol. 12 (1), 126.

Birn, R.M., 2012. The role of physiological noise in resting-state functional connectivity. Neuroimage 62 (2), 864–870.

Birn, R.M., Molloy, E.K., Patriat, R., Parker, T., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V., 2013. The effect of scan length on the reliability of resting-state fmri connectivity estimates. Neuroimage 83, 550–558.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magn. Reson. Med. 34 (4), 537–541.

Chao-Gan, Y., Yu-Feng, Z., 2010. Dparsf: a matlab toolbox for “pipeline” data analysis of resting-state fmri. Front. Syst. Neurosci., 4.

Christoff, K., Gordon, A.M., Smallwood, J., Smith, R., Schooler, J.W., 2009. Experience sampling during fmri reveals default network and executive system contributions to mind wandering. Proc. Natl. Acad. Sci. 106 (21), 8719–8724.

Dutilleul, P., Clifford, P., Richardson, S., Hemon, D., 1993. Modifying the *t*-test for assessing the correlation between two spatial processes. Biometrics, 305–314.

Filevich, E., Lisofsky, N., Becker, M., Butler, O., Lochstet, M., Martensson, J., Wenger, E., Lindenberger, U., Kühn, S., 2017. Day2day: Investigating daily variability of magnetic resonance imaging measures over half a year. BMC Neurosci., (under review).

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci..

Fornito, A., Zalesky, A., Bullmore, E.T., 2010. Network scaling effects in graph analytic studies of human resting-state fmri data. Resting State Brain Act.: Implic. Syst. Neurosci., 40.

Friston, K.J., 2011. Functional and effective connectivity: a review. Brain Connect. 1 (1), 13–36.

Guerra-Carrillo, B., Mackey, A.P., Bunge, S.A., 2014. Resting-state fMRI a window into human brain plasticity. Neuroscientist, (1073858414524442).

Gordon, Evan M., Timothy, O. Laumann, Adeyemo, Babatunde, Huckins, Jeremy F., Kelley, William M., Petersen, Steven E., 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb. Cortex 26.1, 288–303.

Hacker, C.D., Laumann, T.O., Szrama, N.P., Baldassarre, A., Snyder, A.Z., Leuthardt, E.C., Corbetta, M., 2013. Resting state network estimation in individual subjects. Neuroimage 82, 616–633.

Hindriks, R., Adhikari, M., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N., Deco, G., 2015. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fmri? NeuroImage.

Hurlburt, R.T., Alderson-Day, B., Fernyhough, C., Kühn, S., 2015. What goes on in the resting-state? A qualitative glimpse into resting-state experience in the scanner. Front. Psychol., 6.

Lance, C.E., Butts, M.M., Michels, L.C., 2006. The sources of four commonly reported cut-off criteria what did they really say? Organ. Res. Methods 9 (2), 202–220.

Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.-Y., Gilmore, A.W., McDermott, K.B., Nelson, S.M., Dosenbach, N.U., et al., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87 (3), 657–670.

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., et al., 2016. On the stability of bold fmri correlations. Cereb. Cortex.

Lindquist, M.A., Xu, Y., Nebel, M.B., Caffo, B.S., 2014. Evaluating dynamic bivariate correlations in resting-state fmri: a comparison study and a new approach. Neuroimage 101, 531–546.

Logothetis, N.K., 2008. What we can do and what we cannot do with fmri. Nature 453 (7197), 869–878.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412 (6843), 150–157.

Lowe, M., Mock, B., Sorenson, J., 1998. Functional connectivity in single and multislice echoplanar imaging using resting-state fluctuations. Neuroimage 7 (2), 119–132.

Magri, C., Schridde, U., Murayama, Y., Panzeri, S., Logothetis, N.K., 2012. The amplitude and timing of the bold signal reflects the relationship between local field potential power at different frequencies. J. Neurosci. 32 (4), 1395–1407.

Mitra, A., Snyder, A.Z., Blazey, T., Raichle, M.E., 2015. Lag threads organize the brain's intrinsic activity. Proc. Natl. Acad. Sci. 112 (17), E2235–E2244.

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 (3), 586–595.

Nunnally, J.C., 1994. Psychometric Theory. McGraw-Hill, New York.

Ogawa, S., Lee, T.-M., Kay, A.R., Tank, D.W., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc. Natl. Acad. Sci. 87 (24), 9868–9872.

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity mri networks arise from subject motion. Neuroimage 59 (3), 2142–2154.

Prichard, D., Theiler, J., 1994. Generating surrogate data for time series with several simultaneously measured variables. Phys. Rev. Lett. 73 (7), 951.

Rack-Gomer, Anna Leigh, Liau, Joy, Liu, Thomas T., 2009. Caffeine reduces resting-state BOLD functional connectivity in the motor cortex. Neuroimage 46, 1.

Richiardi, Jonas, Eryilmaz, Hamdi, Schwartz, Sophie, Vuilleumier, Patrik, Van De Ville, Dimitri, 2011. Decoding brain states from fMRI connectivity graphs. Neuroimage 56, 2616–2626.

Rosazza, C., Minati, L., 2011. Resting-state brain networks: literature review and clinical applications. Neurol. Sci. 32 (5), 773–785.

Shah, L.M., Cramer, J.A., Ferguson, M.A., Birn, R.M., Anderson, J.S., 2016. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain Behav. 6 (5), e00456.

Shannon, Benjamin John, Dosenbach, Ronny A., Su, Yi, Vlassenko, Andrei G., Larson-Prior, Linda J., Nolan, Tracy S., Snyder, Abraham Z., Raichle, Marcus E., 2013. Morning-evening variation in human brain metabolism and memory circuits. J. Neurophysiol. 109.5, 1444–1456.

Shehzad, Z., Kelly, A.C., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Lee, S.H., Margulies, D.S., Roy, A.K., Biswal, B.B., et al., 2009. The resting brain: unconstrained yet reliable. Cereb. Cortex 19 (10), 2209–2229.

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.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420.

Tagliazucchi, Enzo, Laufs, Helmut, 2014. Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron 82.3, 695–708.

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 (1), 273–289.

Van Dijk, K.R., Hedden, T., Venkataraman, A., Evans, K.C., Lazar, S.W., Buckner, R.L., 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. J. Neurophysiol. 103 (1), 297–321.

Yan, C.-G., Craddock, R.C., Zuo, X.-N., Zang, Y.-F., Milham, M.P., 2013. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. Neuroimage 80, 246–262.

Zou, Q., Miao, X., Liu, D., Wang, D.J., Zhuo, Y., Gao, J.-H., 2015. Reliability comparison of spontaneous brain activities between bold and cbf contrasts in eyes-open and eyes-closed resting states. NeuroImage 121, 91–105.

Zuo, X.-N., Xing, X.-X., 2014. Test-retest reliabilities of resting-state fmri measurements in human brain functional connectomics: a systems neuroscience perspective. Neurosci. *Biobehav. Rev.* 45, 100–118.