NeuroImage 168 (2018) 499–508

Contents lists available at ScienceDirect

NeuroImage

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

Temporal reliability of ultra-high field resting-state MRI for single-subject sensorimotor and language mapping

Paulo Branco[a,b], Daniela Seixas[c,d], São Luís Castro[b,e,*]

[a] Faculty of Medicine, University of Porto, Portugal
[b] Centre for Psychology at University of Porto, Portugal
[c] Dept. of Experimental Biology, Faculty of Medicine, University of Porto, Portugal
[d] Department of Imaging, Centro Hospitalar de Vila Nova de Gaia/Espinho, Vila Nova de Gaia, Portugal
[e] Faculty of Psychology and Educational Sciences, University of Porto, Portugal

ARTICLE INFO

Keywords:
Ultra-high resolution
Resting-state fMRI
Language network
Sensorimotor network
Reliability
Independent component analysis

ABSTRACT

Resting-state fMRI is a well-suited technique to map functional networks in the brain because unlike task-based approaches it requires little collaboration from subjects. This is especially relevant in clinical settings where a number of subjects cannot comply with task demands. Previous studies using conventional scanner fields have shown that resting-state fMRI is able to map functional networks in single subjects, albeit with moderate temporal reliability. Ultra-high resolution (7 T) imaging provides higher signal-to-noise ratio and better spatial resolution and is thus well suited to assess the temporal reliability of mapping results, and to determine if resting-state fMRI can be applied in clinical decision making including preoperative planning. We used resting-state fMRI at ultra-high resolution to examine whether the sensorimotor and language networks are reliable over time — same session and one week after. Resting-state networks were identified for all subjects and sessions with good accuracy. Both networks were well delimited within classical regions of interest. Mapping was temporally reliable at short and medium time-scales as demonstrated by high values of overlap in the same session and one week after for both networks. Results were stable independently of data quality metrics and physiological variables. Taken together, these findings provide strong support for the suitability of ultra-high field resting-state fMRI mapping at the single-subject level.

1. Introduction

Functional magnetic resonance imaging (fMRI) is a key tool for clinical practice because it can be used to map brain function noninvasively. It has been extensively applied in preoperative planning to identify eloquent cortex leading to substantial improvement of clinical outcomes including better risk assessment (Kundu et al., 2013; Petrella et al., 2006) and diminished risk of neurological deficits post-surgery (Hall et al., 2005; Wengenroth et al., 2011). Conventional fMRI mapping protocols for preoperative planning require patients to execute simple tasks in the scanner (task-based fMRI), such as finger-thumb opposition and hand grasping for sensorimotor mapping, and verb-to-noun generation or picture naming for language mapping (for a review see Sunaert (2006)). Evidently, mapping results will depend on the subjects' ability to perform such tasks and this can be an obstacle to the implementation of task-based fMRI in clinical settings, particularly in patients with cognitive or physical impairments (Price et al., 2006).

One straightforward alternative to task-based fMRI is the study of functional connectivity during rest: resting-state fMRI (rs-fMRI). Resting-state fMRI allows the study of spontaneous, low-frequency (< 0.1 Hz) fluctuations that occur at the whole brain. It can be used to segregate functional networks of interest for preoperative planning such as the sensorimotor and language networks (Beckmann et al., 2005; Shirer et al., 2012). Previous studies have demonstrated that rs-fMRI extracted networks are similar to their task-based homologues, both in healthy (Kristo et al., 2014; Mannfolk et al., 2011; Tie et al., 2014) and clinical populations (Branco et al., 2016; Rosazza et al., 2014; Sair et al., 2016). Most importantly, comparisons with the gold-standard direct cortical stimulation indicate that rs-fMRI can efficiently map eloquent cortex (Cocheraeu et al., 2016; Mitchell et al., 2013), as well as contribute significantly to the quality of the mapping procedure when combined with task-based fMRI (Fox et al., 2016).

With promising advances in the field, it is now indispensable to assess if rs-fMRI is temporally reliable enough to be used in clinical decision making. Even though a few studies have reported good group-level temporal reliability of rs-fMRI (Shehzad et al., 2009; Zuo and

* Correspondence to: Faculty of Psychology and Educational Sciences, University of Porto, Rua Alfredo Allen, 4200-135 Porto, Portugal.
E-mail address: slcastro@fpce.up.pt (S.L. Castro).

http://dx.doi.org/10.1016/j.neuroimage.2016.11.029
Accepted 12 November 2016
Available online 15 November 2016

Xing, 2014), studies on single subjects are scarce and somewhat inconsistent. For example, Mannfolk et al. (2011) and Meindl et al. (2010) have observed good temporal reliability of the sensorimotor and default mode networks, respectively. Pinter et al. (2016) have shown that several resting-state networks, including the sensorimotor, visual and default mode networks, have excellent within region-of-interest (ROI) reliability, but poor (at best) reliability when examining all the voxels included in group-level networks. Perhaps more strikingly, Kristo et al. (2014) performed a comprehensive test-retest comparison of rs-fMRI and task-based fMRI for single-subject sensorimotor mapping. They were able to consistently identify the sensorimotor network with rs-fMRI, but its reliability was poor. They also found that task-based fMRI was significantly more reliable than rs-fMRI, especially with conservative statistical thresholds (i.e., masks of 5000 voxels or less). So, in order to ascertain the robustness of resting-state mapping more studies are needed on single-subject test-retest metrics and which factors subtend its reliability.

An important confounding factor in fMRI test-retest studies is the relative poor signal-to-noise ratio (SNR) at conventional scanner fields of, e.g., 3 T. Poor SNR requires image acquisition with spatial resolution between 3 to 4 mm^3 and post-processing procedures such as spatial smoothing that hinder the anatomical precision necessary for accurate assessment of temporal reliability. The lack of high spatial resolution may result in poor estimation of the true reliability of rs-fMRI. This issue can be addressed thanks to the increasing availability of ultra-high field MRI scanners at 7 T; these have millimetre spatial resolution and are better able delineate functional networks (De Martino et al., 2011; Hale et al., 2010; Sanchez-Panchuelo et al., 2010). Given its excellent spatial resolution and high SNR, ultra-high field MRI is well suited to investigate temporal reliability, and it can provide compelling evidence as to whether resting-state fMRI is feasible, or not, in clinical settings.

In this study, we use ultra-high resolution 7 T MRI to examine if rs-fMRI can consistently identify two functional networks of major importance in preoperative planning – the sensorimotor and language networks. We will assess the temporal reliability of mapping outcomes in the same session (intrasession reliability) and after one week (intersession reliability). Finally, we will test whether data-quality and physiological measures affect the reliability of this mapping procedure.

2. Methods

2.1. Participants

Data used in this study were obtained from a publicly released dataset from the Consortium for Reliability and Reproducibility (CoRR, Zuo et al., 2014) project (Gorgolewski et al., 2015). This dataset contains test-retest data from 22 adult volunteers scanned four times over a time-period of one week. Two subjects were excluded from our sample: subject 5 because he was left-handed, and subject 11 because he was scanned with a different voxel size than the remaining subjects. Thus the dataset we used consisted of 20 right-handed subjects (10 men+10 women) that were scanned with the same rs-fMRI sequence. Their mean age was 24.8 ± 1.9 years.

2.2. Imaging protocol

The imaging protocol is presented by Gorgolewski et al. (2015); a brief description follows. Subjects were scanned in a Siemens 7 T Magnetom scanner with a combined birdcage transmit and 24-channel phased array receiving coil. Resting-state fMRI data was acquired using a 2D sequence with the following parameters: time of repetition (TR) =3000 ms; time of echo (TE)=17 ms; partial Fourier 6/8; GRAPPA acceleration factor iPAT=3; field of view (FOV)=192×192 mm^2; flip angle 70°; slice thickness=1.5 mm; in-plane pixel size=1.5×1.5 mm^2; and axial slices=70. Three hundred functional volumes were obtained in 15 minutes. A high-resolution 3D MP2RAGE image was also acquired for each subject, with the following parameters: TR=5.0 s; TE=2.45 ms; partial Fourier 6/8; GRAPPA acceleration factor iPAT=2; FOV=225×224×168 mm^2; slice thickness=0.75 mm; in-plane pixel size=0.75 × 0.75 mm^2, and axial slices=70.

2.3. Procedure

As before, the procedure is detailed in Gorgolewski et al. (2015) and here we only recapitulate the main points. Subjects had two scanning sessions exactly one week apart, and in each session two rs-fMRI protocols were performed within a one-hour interval (15 min each, 300 volumes). In the present study, we examined intrasession reliability by comparing protocols 1 and 2 (also referred to as time 1 and time 2; same scanning session), and intersession reliability by comparing protocols 1 and 3 (time 1 vs. time 3; one week interval).

Subjects were instructed to relax (but not sleep) in the scanner with their eyes open, while viewing a fixation cross presented on the screen. Before each scanning session, physiological and behavioural data were collected. These included mood, sustained attention, blood pressure (diastolic and systolic, left and right), pulse (left and right), hydration, caffeine intake and sleep habits.

2.4. Data preprocessing

Data preprocessing and subsequent analyses were performed using the Oxford Centre for Functional Magnetic Resonance Imaging of the Brain Software Library (FMRIB, Oxford U.K.; FSL version 5.0.9). Structural data was skull-stripped using the procedure as follows. First, the second inversion image (INV2) was multiplied by the uniform contrast image (UNI) to reduce noise amplification outside the brain. Then, the resulting image was skull-stripped using BET (Smith, 2002) with the built-in bias-field and neck removal option. Finally, this image was binarized, refined manually, and used as a mask to extract the brain from the original UNI image. Cerebral spinal fluid (CSF) and white-matter (WM) masks were obtained from the skull-stripped brain image using FAST (Zhang et al., 2001). To ensure no overlap between gray matter (GM) and CSF masks, each subject's CSF mask was manually edited to include only voxels within the lateral, third and fourth ventricles. Finally, CSF and WM masks were conservatively eroded by 1 voxel.

For resting-state data, the initial three volumes were discarded to account for T1 saturation effects. Motion correction was performed by aligning all volumes to a middle reference using MCFLIRT (Jenkinson et al., 2002) and corrected for magnetic susceptibility artefacts using FUGUE (Jenkinson, 2003). Data were skull-stripped using BET (Smith, 2002) and spatially smoothed using a Gaussian full width at half maximum (FWHM) filter of 3 mm; smoothing was performed so that other pre- and post-processing steps were viable (e.g., ICA denoising), but was kept at a minimum to preserve high spatial resolution. Data were denoised using ICA-AROMA (Pruim et al., 2015a). White-matter and CSF mean time series were extracted by computing the average time courses within structural WM and CSF masks, and removed from the data through multiple linear regression. Finally, residuals were high-pass filtered at a FWHM 100s cutoff.

To reduce the impact of coregistration errors and avoid additional smoothing of the data, all analyses were performed on the subjects' functional space. To compare across time-point protocols, functional maps were warped to each subject's structural image and kept at the native resolution of 1.5 mm^3 isotropic voxels using FLIRT (Jenkinson et al., 2002) with boundary-based registration (Greve and Fischl, 2009). After coregistration, we applied spatial normalization to warp the standard-space masks into each subject's structural space and to perform group-level analyses. Transformation matrices from structural to standard MNI space were obtained through 12 degrees-of-freedom

registration with FLIRT, and were further refined by non-linear registration with FNIRT (Andersson et al., 2007).

2.5. Data analysis

Resting-state networks were extracted with independent component analysis (ICA) followed by a dual regression approach. To do so, preprocessed data was fed into MELODIC single-session ICA (Beckmann and Smith, 2004). The number of extracted components was estimated with FSL default Laplacian approximation (Beckmann and Smith 2004). Independent components (ICs) were thresholded using a mixture-model cut-off of 0.5 for an equal weight on false-positives and false negatives (Woolrich et al., 2005). All analyses were performed at the single-subject level. Although each subject performed multiple resting-state protocols, each run was analysed separately to mirror mapping results as if they had been acquired only once, as it typically happens in clinical settings.

Sensorimotor and language ICs were identified using a template-matching procedure similar to that used in previous studies (Branco et al., 2016; DeSalvo et al., 2016; Tie et al., 2014). Briefly, the ICs were identified through spatial comparison against previously published group templates (Shirer et al., 2012). To do this, we calculated spatial overlap using the Dice coefficient with the following equation:

Eq. (1). Dice-coefficient

$$Dice = \frac{2 \times M_{overlap}}{(M_1 + M_2)} \quad (1)$$

where M_1 and M_2 represent the number of supra-threshold in each mask, and $M_{overlap}$ the number of supra-threshold voxels that are present in both masks. The Dice coefficient provides a measure of overlap between two masks, M_1 and M_2; it varies between 0 and 1, higher values representing higher similarity. Dice coefficients were computed between each IC and the sensorimotor and language group templates, and were ranked according to spatial similarity, from highest to lowest. The five highest ranked ICs for each network were visually inspected by two of the authors (PB and DS) with blinding to the classification ranking. Each IC was independently rated using a seven-point scale according to the confidence with which it was attributed to each network (1=very unlikely; 7=very likely). Independent components with average confidence ratings of 6 and above were considered as belonging to the corresponding network. These ICs were merged into a single mask for each network after applying a threshold of z=5. These masks were then used to re-estimate the corresponding network using a dual-regression approach (Beckmann et al., 2009; Filippini et al., 2009). First, for each subject, the mask containing the selected ICs was spatially regressed against the whole-brain preprocessed data to estimate each network's timecourse. Next, the estimated timecourses were temporally regressed against the whole-brain data to obtain subject-specific networks. Finally, resulting maps were converted to z scores. With this approach we expected to improve the estimation of single-subject networks by solving potential ICA overdecomposition and obtaining whole-brain connectivity indices that were appropriate for the interpretation of reliability metrics.

2.6. Group level results

To obtain group-level maps for each network, results from the three time-point protocols were averaged for each subject and network. These averages were used in a nonparametric one-sample t-test using the FSL randomise permutation tool (Winkler et al., 2014) with a voxelwise FWE-corrected threshold of $p < .01$.

2.7. Reliability metrics

Temporal reliability of resting-state language and sensorimotor networks was estimated with the two most widely used fMRI reliability metrics: Dice coefficient (Rombouts et al., 1998) and intraclass correlation (ICC, Shrout and Fleiss, 1979).

The Dice coefficient quantifies the proportion of supra-threshold voxels that are common across measurements, and thus it can vary according to the threshold used. For this reason, we calculated Dice coefficients at fixed z thresholds, from z=2 to z=20 in steps of 0.5. Additionally, as thresholds can vary across subjects and sessions (Gorgolewski et al., 2012), we also thresholded each subject's mask such that a fixed number of significant voxels were kept within a mask, ranging from 1000 to 20,000 voxels in steps of 1000, and computed corresponding Dice coefficients (see Kristo et al. (2014) for a similar approach). Networks were in native (ultra-high) resolution with isotropic voxel sizes of 1.5 mm^3. The total volume of the masks ranged from 3375 to 67,500 mm^3 (1000 to 20,000 voxels, respectively).

Intraclass correlation examines the stability of the activity at all brain voxels with no thresholding required. It takes into account the magnitude of activation and not just its spatial extent. As in Kristo et al. (2014) and Raemaekers et al. (2007), we used the two-way random ICC (formula 2.1 from Shrout and Fleiss, 1979) for absolute agreement between pairs of measures. This ICC variant examines the proportion of total variance that is explained by intra-voxel variance and provides an individual measure of the stability of voxel intensities for pairs of measures. Although it has a few known limitations (for example, pre-processing steps such as spatial smoothing can bias the ICC towards higher reliability; Bennett and Miller, 2010; Caceres et al., 2009), we used here because it is well suited to data modeling and it would allow us to compare our results with those of Kristo et al. (2014), who also examined test-retest reliability in single-subject resting-state mapping. The following equation was used to calculate the ICC:

Eq. (2). Two-way intraclass correlation for absolute agreement

$$ICC_{within} = \frac{BMS - EMS}{BMS + (k-1)EMS + k(JMS - EMS)/n} \quad (2)$$

where BMS refers to the mean square of the variance in z values between voxels, JMS to the mean square of the column differences in voxel z values between measurements, EMS to the mean squared error term, k the number of sessions (2) and n to the number of targets (# of voxels). Intraclass correlation values were transformed using a Fisher r to z transformation to approach a normal distribution before parametric statistics:

Eq. (3). Fisher r-to-z transformation

$$ICC' = \left(\frac{1}{2}\right)\log\left(\frac{1+ICC}{1-ICC}\right) \quad (3)$$

For both Dice and ICC reliability metrics, we performed analyses at the whole-brain level and analyses restricted to regions of interest (ROI). We selected different ROIs from those used in the ICA template-matching procedure to prevent selection bias. For sensorimotor regions, we used a combined mask from the precentral and postcentral gyrus taken from the Harvard-Oxford cortical atlas (Desikan et al., 2006), with a minimal probability of 20%. For language regions, we used Fedorenko et al. (2010) ROIs combined into a single mask; these ROIs were validated in healthy subjects and represent high-level language regions that are consistently identified across large samples of individuals (Mahowald and Fedorenko, 2016).

For an additional measure of mapping efficiency, we examined the percentage of supra-threshold voxels that lay within ROIs using masks sizes of 1000, 5000 and 15000 voxels. This measure—a trade-off between sensitivity and specificity—indexes whether the networks were reliably confined within classical ROIs.

Unless otherwise specified, the reliability metrics described above were analyzed using repeated-measures ANOVAs with Network (sensorimotor *vs.* language), WB/ROI (whole-brain *vs.* ROI values) and Comparison Type (intrasession *vs.* intersession) as within-subject factors. Post hoc comparisons were performed with paired sample t-

tests, with Bonferroni correction for multiple comparisons. Significant results are reported in the text; full statistical results can be consulted in the Supplementary material, Tables s2 to s8.

2.8. Potential confounding factors

We examined whether the temporal reliability of rs-fMRI could be explained by measurement related factors. We selected three indices that have been singled out as critical for test-retest reliability: temporal SNR (Bennett and Miller, 2010), head motion (Gorgolewski et al., 2013) and between-protocol coregistration error (Fernandez et al., 2003). Temporal SNR was calculated after motion correction and immediately before spatial smoothing, after removing the first three functional volumes to exclude T1 saturation effects (as in Gorgolewski et al. (2013)). We computed it for each subject and protocol by calculating the mean and standard deviation of the data (across all brain voxels) and taking their ratio. Temporal SNR was averaged across protocols 1 and 2 to obtain intrasession temporal SNR estimates, and across protocols 1 and 3 to obtain intersession estimates. Head movement was indexed by the average framewise displacement across all volumes (Power et al., 2014) in order to provide a frame-by-frame quantification of head movement. Estimates for intrasession head motion were obtained by averaging protocols 1 and 2, and intersession head motion by averaging protocols 1 and 3. Coregistration error was assessed by calculating the spatial correlation between the mean functional images of each protocol after warping from functional to structural space (Gorgolewski et al., 2013). In order to facilitate the interpretation of coregistration error, we computed dissimilarity scores by subtracting the correlation coefficient from 1 ($1-r$), higher values representing more error. Intra- and intersession coregistration errors were computed by comparing protocols 1 vs. 2, and 1 vs. 3, respectively.

We also examined how other measures available in the dataset might have affected the imaging results, albeit in an exploratory way. We chose a set of variables based on their statistical distribution (normality), variability, and how well they stood for each subject-related category (sleep habits, hydration and cardiovascular data). For sleep habits, we calculated the difference between the hours slept in the previous night and the usual average sleeping hours (self-report). For hydration, we used the thirst measure in which subjects reported how hydrated they felt on a scale from 1 to 9. For cardiovascular data, we derived two measures: pulse, by averaging the left and right arm pulse counts; and blood pressure, by averaging the systolic blood pressure of both arms. As these measures were collected only once per scanning session (see Procedure), we used data from the first session for the intrasession reliability analysis, and the average of both sessions (day one, and one week after) for the intersession reliability analysis.

To estimate how each of the selected variables impacted on the reliability estimates, multiple regressions were calculated (as in Gorgolewski et al. (2013)). Given the relatively low ratio of subjects to variables, two separate models were fit to the data: the first model included data-quality metrics (temporal SNR, head motion and coregistration error), and the second included subject characterization data (sleep habits, hydration, pulse and systolic blood pressure). To further examine the relative importance of each variable, we used the Lindeman-Merenda-Gold metric with the R package relaimpo (Grömping, 2006); this metric iteratively reorders the variables to check for the contribution of each variable to the total variance explained by the model. The models were computed for intrasession and intersession Dice and ICC reliability estimates of the sensorimotor and language networks.

3. Results

3.1. Extraction and classification of the sensorimotor and language networks

The average number of components extracted with the ICA procedure was similar across the three time-point protocols: 49.9, 49.4 and 52.3 for protocols 1, 2 and 3, respectively ($F(2, 38)=1.1$, $p=.34$, *ns*). Sensorimotor networks were identified for all subjects in the three protocols with high interrater agreement (Krippendorf $\alpha=.86$). For each subject and protocol, there was an average of 2.2 sensorimotor ICs (range 1–4). Regarding the language networks, there were 3 cases out of 60 (20 subjects×3 protocols) in which no IC received an average confidence rating of 6 or 7 (see Methods). Upon visual inspection of these cases, we decided to include the next best IC candidate, which indeed had confidence ratings of 5 or higher and was rated 6 by one of the raters. For each subject and protocol, an average of 1.09 ICs (range 1–2) were selected for the language networks; interrater agreement was moderate (Krippendorf $\alpha=.72$).

3.2. Group-level sensorimotor and language networks

The regions identified as the sensorimotor and language networks (see Fig. 1) were consistent with previous studies. For the sensorimotor network, we found significant clusters in the bilateral precentral and postcentral gyrus, as well as in the supplementary motor area. For the language network, the significant clusters were in classical language regions including the left inferior frontal gyrus and the left posterior middle temporal gyrus. Table 1 shows a summary of the three largest clusters for each network and respective MNI coordinates. Sensorimotor and language group maps can be further inspected at Neurovault.org (http://neurovault.org/collections/CIQVCXUR).

3.3. Temporal reliability estimated with Dice overlap at fixed z values

High temporal reliability was observed using Dice overlap values (see Fig. 2a for Dice values at different threshold levels). In whole-brain analyses, intrasession Dice values for the sensorimotor network were 0.632, 0.595 and 0.562, at thresholds of $z=3$, $z=6$, and $z=9$, respectively. Corresponding values for intersession Dice were 0.599, 0.584, and 0.532. High Dice values were also observed for the language network: intrasession 0.593, 0.560, 0.532 and intersession 0.580, 0.526 and 0.488, at thresholds of $z=3$, $z=6$ and $z=9$, respectively. Dice values in ROI analyses were generally higher than in the whole brain ($Fs(1, 19) > 237.85$, $ps < .001$, $n_p^2 > .93$): in the sensorimotor network they reached 0.890, 0.849 and 0.805 intrasession, and

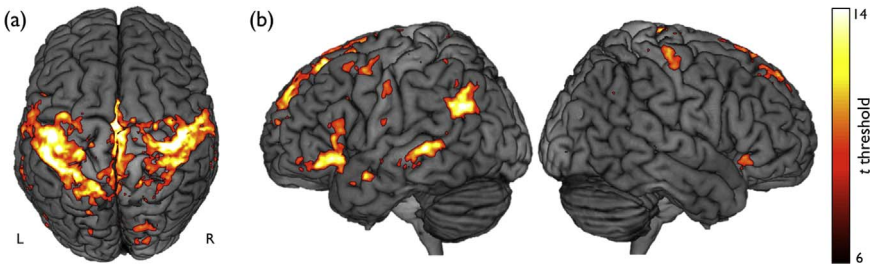

Fig. 1. Group-level statistic maps for the sensorimotor (a) and language networks (b) obtained with non-parametric one-sample *t*-tests (voxelwise threshold at $p < .01$, FWE corrected).

Table 1
Group-level results for the sensorimotor and language networks (three largest clusters).

Brain region	# Voxels	MNI coordinates			Z score
		x	y	z	
Sensorimotor network					
L Postcentral Gyrus	21119	−50	−26	56	24.9
L Precentral Gyrus		−46	−15	57	21.7
L Postcentral Gyrus		−49	−31	56	20.8
L Postcentral Gyrus		−40	−26	58	20.1
L Postcentral Gyrus		−48	−33	58	19.7
L Precentral Gyrus		−37	−21	47	19.2
R Precentral Gyrus	13347	38	−24	66	24.9
R Precentral Gyrus		34	−23	69	23.6
R Postcentral Gyrus		45	−17	55	21.2
R Postcentral Gyrus		42	−23	64	19.9
R Postcentral Gyrus		44	−20	60	18.1
R Postcentral Gyrus		43	−20	62	18
L Juxtapositional Lobule Cortex (SMA)	11582	−2	−11	56	23.8
R Precentral Gyrus		1	−29	77	22.9
L Juxtapositional Lobule Cortex (SMA)		−2	−7	54	21.1
R Precentral Gyrus		2	−13	77	19.4
LR Precentral Gyrus		0	−25	78	19.2
R Precentral Gyrus		3	−19	76	18.4
Language network					
L Superior Frontal Gyrus	29742	−10	32	56	18
L Paracingulate Gyrus		−3	57	15	16.5
R Precentral Gyrus		4	−20	76	16.3
L Frontal Pole		−15	56	25	16
L Frontal Pole		−11	53	42	15.9
R Postcentral Gyrus		8	−33	76	15.8
L Posterior Middle Temporal Gyrus	15118	−48	−34	−1	18.7
L Posterior Middle Temporal Gyrus		−49	−31	−4	18.5
L Posterior Middle Temporal Gyrus		−66	−34	1	18.2
L Superior Middle Temporal Gyrus		−55	−37	2	18.2
L Angular Gyrus		−58	−53	30	17.2
L Angular Gyrus		−48	−59	29	17.1
L Frontal Orbital Cortex	7433	−48	26	−6	16.5
L Frontal Pole		−53	35	−10	16.3
L Frontal Orbital Cortex		−46	21	−6	15.8
L Frontal Orbital Cortex		−47	28	−8	15.8
L Inferior Frontal Gyrus, Pars Triangularis		−51	21	−6	15.3
L Frontal Orbital Cortex		−44	29	−5	15.1

Note: Results listed in the table were obtained after a $z > 10$ (FWE-corrected) threshold for the sensorimotor network, and a $z > 8$ (FWE-corrected) threshold for the language network. Three largest clusters ordered by number voxels for each cluster (# voxels), and a maximum of 6 peak coordinates per cluster in MNI space. R, Right. L, Left. Labels taken from the Oxford-Harvard Structural Cortical Atlas.

0.867, 0.817 and 0.769 intersession, at z thresholds of 3, 6 and 9, respectively. Corresponding values for the language network were 0.744, 0.712 and 0.689 intrasession, and 0.736, 0.695 and 0.659 intersession.

At the three threshold levels reported above, the differences between intrasession and intersession Dice values were not significant (Fs(1, 19) < 2.96, *ps* > .10, *ns*; all interactions *F*s(1, 19) < 1.95, *all ps* > .18). However, Dice values were higher in the sensorimotor than in the language network (main effects of network, *F*s(1, 19) > 11.26, *ps* < .01, n^2_p > .37) and this effect was more pronounced within ROIs than in the whole brain (interactions network x ROI/WB, *F*s(1, 19) > 63.33, *ps*n^2_p < .001, > .77).

Despite high Dice coefficients, the volume of the maps varied considerably between sessions and subjects according to mask size (see Fig. 2b). As larger masks lead to higher coefficients (Kristo et al., 2014), in order to have reliability estimates unbiased by mask size we computed Dice overlap using fixed size-based masks.

3.4. Dice overlap with fixed size-based masks

An alternative thresholding procedure is to select a z cutoff value such that for a given statistical significance level all masks have a fixed number of active voxels (Kristo et al., 2014). We evaluated the overlap between protocols at a range between 1000 and 20,000 voxels in steps of 1000 voxels. At the smallest mask size (1000 voxels) the average z threshold for the three protocols was 16.16 ± 2.39 for the sensorimotor and 14.49 ± 1.50 for the language network. At the largest size (20,000 voxels), corresponding values were 8.93 ± 2.43 (sensorimotor network) and 7.64 ± 1.57 (language network). Following visual inspection of the differently sized masks, we defined 5000 voxels as a conservative threshold and 15,000 voxels as a conventional threshold. For clarity, only the results for these thresholds will be reported in the text; values of the full range of mask sizes are shown in Fig. 3.

Again, Dice values were high. In whole-brain analyses, intrasession Dice values for the sensorimotor network reached 0.574 and 0.602 at conservative and conventional thresholds, respectively; corresponding values for intersession reliability were 0.548 and 0.588; for language, analogous values for intrasession reliability were 0.518 and 0.557, and for intersession reliability 0.512 and 0.541 (conservative and conventional thresholds, respectively). In the ROI analyses, Dice values were generally higher (all *ps* < .001, see Fig. 3), but also not significantly different within or between sessions (*F*s < 1, *ns*). Specifically, for the sensorimotor network, intrasession Dice values were 0.693 and 0.616 and intersession values were 0.675 and 0.583, at conservative and conventional thresholds, respectively; for the language network, corresponding Dice values were 0.632 and 0.559, intrasession, and 0.617 and 0.549, intersession, respectively. With conventional thresholds, the sensorimotor networks were more reliable than the language networks (main effect of network, $F(1, 19)=6.29$, $p=.02$, $n^2_p=.25$), but with conservative thresholds this effect did not reach significance ($F(1, 19)=3.68$, $p=.07$, *ns*).

3.5. Temporal reliability estimated with intraclass correlation

In whole-brain analyses, good ICC values were observed for the somatosensory network: intrasession ICC was 0.65 ± 0.11 (range 0.48–0.83), and intersession ICC 0.61 ± 0.14 (range 0.3–0.92). The results were similar for the language network: intrasession ICC: 0.63 ± 0.12 (range 0.37–0.77), intersession ICC: 0.60 ± 0.14 (range 0.35–0.96; see Fig. 4). Within ROIs, ICCs were higher than in the whole-brain ($F(1, 19)=97.3$, $p < .001$, $n^2_p=.84$; interaction comparison type x ROI/WB, $F(1, 19)=3.51$, $p=.08$, *ns*; all remaining interactions, $F < 1$). They reached 0.75 ± 0.13 (range 0.39–0.91) intrasession and 0.70 ± 0.15 (range 0.32–0.93) intersession in the somatosensory network. Likewise for the language network: 0.73 ± 0.11 (range 0.48–0.86) intrasession, and 0.69 ± 0.12 (range 0.53–0.97) intersession. The effects of comparison type and of network were not significant, both $F < 1$.

3.6. Resting-state mapping specificity

To better probe the reliability of mappings across sessions, we inspected the percentage of voxels within ROIs at mask sizes of 15000, 5000 and 1000 voxels. The sensorimotor network was well delimited within the ROIs, with average values for the three protocols ranging from 58% to 62% at the conventional 15000 voxel mask, 72% to 78% at the conservative 5000 voxel mask, and 84% to 89% voxels at the smallest mask (1000 voxels). Language networks were not so well delimited within the ROIs, with 43% to 46% of all voxels falling inside the conventional mask, 55% to 58% in the conservative one, and 66% to 69% in the smallest one. ANOVAs confirmed that the sensorimotor network was better captured in the ROIs than the language network in all mask sizes (all *ps* < .001).

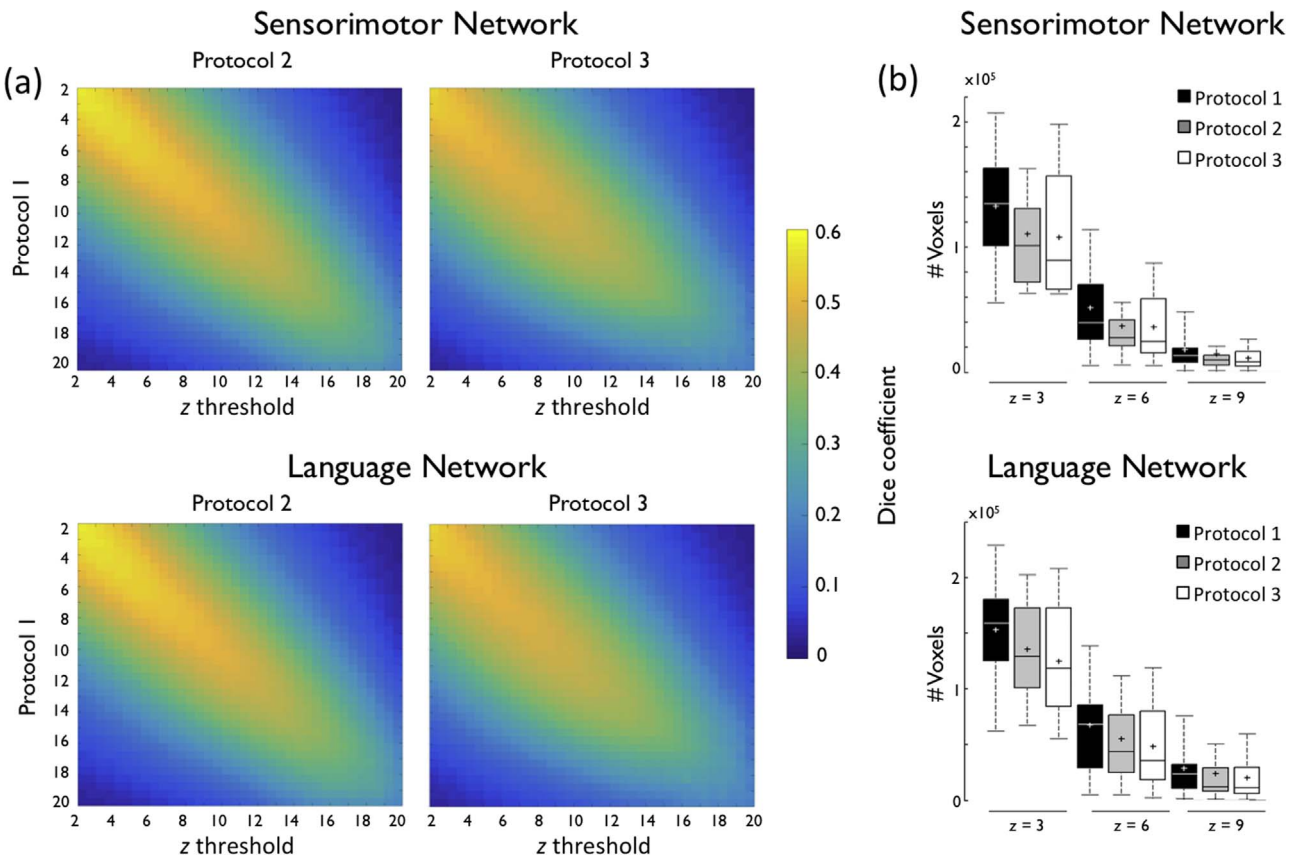

Fig. 2. (a) Mean Dice coefficient for sensorimotor (up) and language (bottom) networks for intrasession (protocol 1 vs. protocol 2) and intersession (protocol 1 vs. protocol 3) at different thresholds, from z=2 to z=20. (b) Boxplots with average map sizes for the sensorimotor (up) and language networks (bottom), and for protocols 1 to 3, at thresholds of z=3, z=6 and z=9.

3.7. Reliability, data quality and subject related factors

As mentioned before, two regression analyses were calculated: the first on data quality metrics (temporal SNR, head motion and coregistration error) and the second on subject-related confounds pertaining to sleep (relative hours of sleep in the previous nights), hydration (relative water intake), and cardiovascular parameters (pulse and systolic blood pressure). Both analyses were tested independently on two reliability metrics: whole-brain Dice at a fixed mask size of 15000 voxels and whole-brain ICC.

Regarding data quality in relation to intrasession reliability, no significant relationships were observed with Dice values (sensorimotor network, $F(3, 16)=1.5$, $p=.26$, *ns*; language network, $F < 1$, *ns*). Fitting the same model to ICC also led to non-significant results (sensorimotor and language networks, $Fs < 1$, *ns*). Considering intersession reliability, data quality factors significantly explained the Dice values in the sensorimotor network ($R^2=44\%$, adjusted $R^2=34\%$, $F(3, 16)=4.19$, $p=.023$), but not in the language network ($F(3, 16)=1.43$, $p=.27$, *ns*). Only coregistration error contributed significantly to the model ($t=3.09$, $p=.007$) with a relative contribution of 33% (head motion: 6.4%, $p=.25$; temporal SNR: 5.3%, $p=.80$). The same model applied to the ICC metric revealed a non-significant R^2 of 22% (sensorimotor network: $F(3, 16)=1.53$, $p=.25$, *ns;* language network: $F(3, 16)=1.1$, $p=.37$, *ns*).

Regarding the analyses on subject related factors, no significant results emerged neither in intrasession nor in intersession reliabilities in Dice or ICC metrics of the sensorimotor and the language networks (all $Fs < 1$; for details, see Supplementary Material, Tables s5 to s8).

3.8. Additional examination of data quality metrics and temporal reliability

Coregistration error significantly explained the Dice reliability metrics in the intersession comparison, but not in the intrasession one. To examine if this result was a consequence of repositioning the head inside the scanner, we ran additional analyses including the second protocol of the second session (see Procedure). This fourth protocol, hereafter protocol 4, differed from the second protocol of the first session in that the subject exited the scanner, was asked to sit upright and was moved into the scanner again (the subject remained in the scanner between protocols 1 and 2). Thus, head position had to be reset between same-day protocols 3 and 4. The same pre- and post-processing pipeline as applied in the previous analyses was used, as well as same multiple-regression modeling approach (see Methods). Protocols 3 and 4 were then compared using Dice coefficients at a fixed mask-size of 15000 voxels and whole-brain ICCs[1].

When comparing coregistration error between protocols 3 and 4, one subject stood out as an outlier (subject 18, 3.4 SDs from the mean). This subject was excluded from further analyses due to the violation of normality in the sample (Kolmogorov-Smirnov test, $p < .05$). A repeated-measures ANOVA on coregistration error with Comparison Type (protocols 1–2 or intrasession 1, protocols 1–3 or intersession, and protocols 3–4 or intrasession 2) as within-subject factor showed significant differences between comparisons ($F(2, 36)=26.77$, $p < .001$, $n_p^2=.60$). Coregistration error between protocols 1 and 2 (0.004 ± 0.001) was significantly lower than between protocols 1 and 3 (0.01

[1] Average Dice coefficients (15000 voxels) were 0.527 for the sensorimotor and 0.485 for the language network. Average ICC were 0.545 for the sensorimotor and 0.525 for the language network.

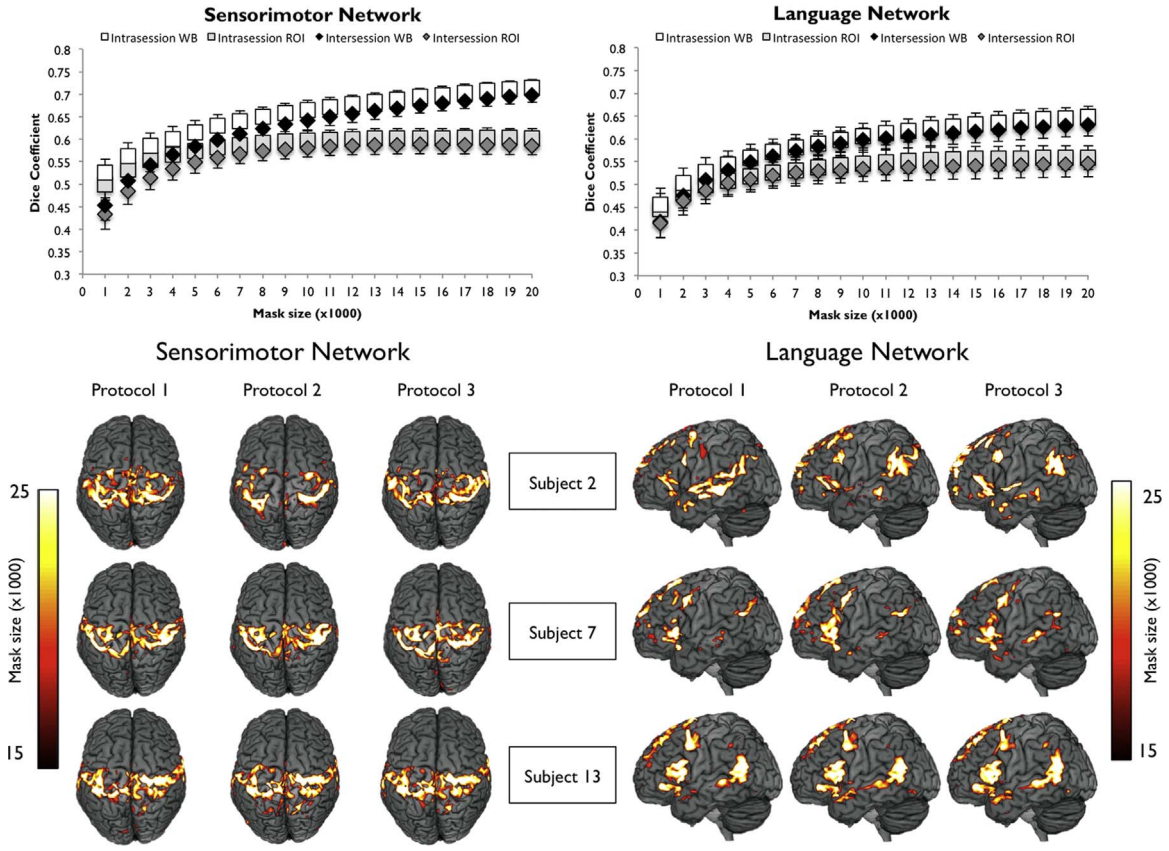

Fig. 3. Upper panel: Whole-brain (WB) and ROI mean Dice coefficients, intrasession and intersession, for the sensorimotor (left) and language networks (right), at fixed map sizes from 1000 to 20,000 voxels in steps of 1000. Vertical bars show standard errors. Bottom panel: Maps from three representative subjects [S2, S7, S13] at each time point (protocols 1, 2 and 3), with *z* thresholds set at a fixed map size of 15,000 voxels. Images presented in standard MNI space for illustration purposes.

± 0.003; $p < .001$) and protocols 3 and 4 (0.006 ± 0.003; p=.018), and the difference between these two was also significant (lower coregistration error for intrasession 2 than for intersession, p=.018). The data quality model applied to the Dice coefficients revealed non-significant R^2s of 21% for the sensorimotor network ($F(3, 15)=1.3$, p=.31, *ns*) and 11% for the language network ($F < 1$, *ns*). The same pattern emerged for the ICC reliability metric, with non-significant R^2s of 14% for the sensorimotor network ($F < 1$) and 27% for the language network ($F(3, 15)=1.84$, p=.18, *ns*). Full statistical data regarding these analyses can be consulted in the Supplementary material, Table s9.

4. Discussion

The aim of this study was to determine the temporal reliability of

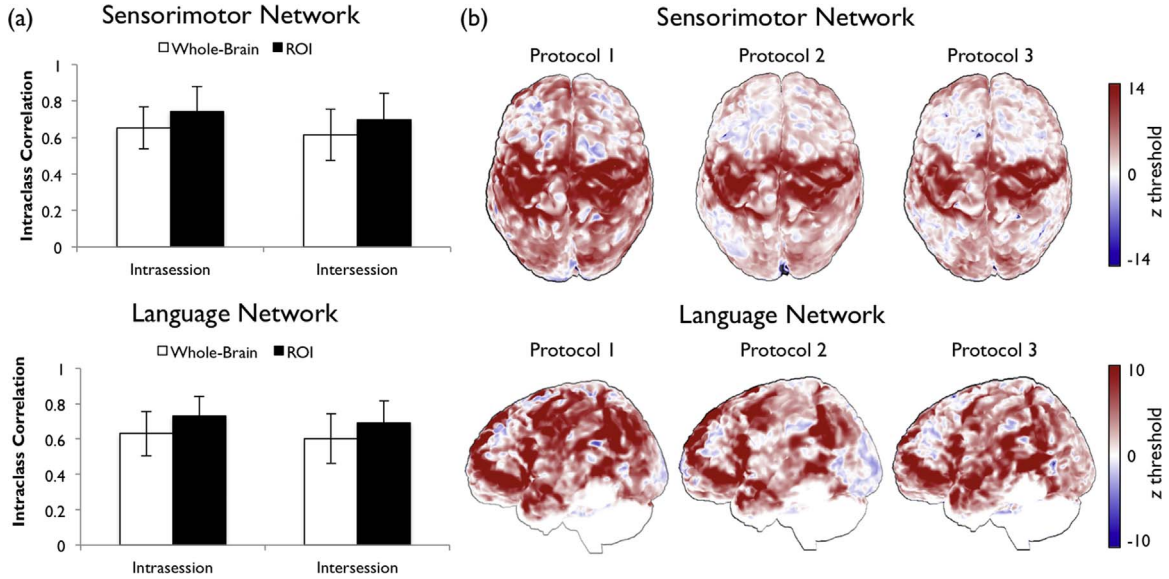

Fig. 4. (a) Whole brain and ROI mean intraclass correlation values for the sensorimotor and language networks for whole-brain and within ROIs, for intrasession and intersession. Vertical bars show standard deviations. (b) Unthresholded maps for the sensorimotor (top) and language (bottom) networks of a representative subject [S6, average whole-brain ICC 0.63]. Images presented in standard MNI space for illustration purposes.

resting-state fMRI at ultra-high resolution for single-subject mapping of the sensorimotor and language networks. We examined reliability in two different time-scales: within one hour or so of the same recording session (intrasession reliability) and in a different session one week later (intersession reliability). Our results can be summarized in three major points. First, both the language and the sensorimotor networks were accurately identified in each subject in the three time points; this reveals the suitability of rs-fMRI as a tool for single-subject brain mapping. Second, ultra-high resolution rs-fMRI mapping was stable over time, with robust reliability metrics within as well as between sessions. Third, neither data quality metrics nor subject related physiological variables affected the reliability of the language network mappings, but coregistration error lead to lower Dice coefficients in the sensorimotor network. Taken together, these findings indicate that sensorimotor and language mapping via ultra-high resolution rs-fMRI is robust enough to be used at the single-subject level and thus holds promise for clinical applications.

The extraction of functional networks on the basis of individual data, including patients with atypical brain morphology, is one of the major challenges to the suitability of rs-fMRI for single-subject brain mapping. We used ICA as a tool for that purpose because previous studies had shown that it was feasible even in subjects with altered brain anatomy (Branco et al., 2016; Rosazza et al., 2014; Sair et al., 2016). We found that sensorimotor and language networks were consistently identified by two blinded raters with high (sensorimotor) and moderate (language) interrater agreement. ICA overdecompositions were observed for the sensorimotor network (2.2 components on average), but not for the language network (1.1 components on average). The potential problem of overdecomposition was addressed by employing a dual regression approach that is well established in the field (Beckmann et al., 2009; Filippini et al., 2009), but unlike in most studies we used masks extracted at the single-subject level (e.g., Camchong et al., 2009) instead of group-level masks. This innovation proved to be effective, as the identified networks were able to capture well-known brain regions for sensorimotor and language functions delimited within typical ROIs. An alternative approach would have been to limit the ICA dimensionality to a reduced number of ICs (DeSalvo et al., 2016). However, this has drawbacks such as the subjectivity associated with choosing the adequate number of ICs and the risk of merging the principal network with unrelated ICs or noise, that we think might undermine its use for single-subject mapping. The high reliability results that we have obtained lend credence to the adequacy of our procedure.

Regarding the temporal stability of the rs-fMRI mappings, the different metrics used to estimate reliability converged on very positive results. The intra- and intersession Dice coefficients obtained in the whole-brain were quite high (about 0.57) and even higher within ROIs (about 0.65). Note that when applying a fixed *z* threshold, the number of active voxels varied considerably across subjects and time-point protocols, a consequence of the well-known difficulty in setting an adequate threshold for single-subject data (Gorgolewski et al., 2012). As reliability scores are expected to increase at larger mask sizes (Kristo et al., 2014), applying the same threshold criteria irrespective of mask size to different subjects may lead to over- or underestimation of Dice scores. This led us to pursue an alternative thresholding procedure such that the reliability estimates would be comparable across subjects and time point protocols. We used the same approach as Kristo et al. (2014) in which the threshold is set according to the number of significant active voxels. Even though this is a rather conservative criterion, we obtained Dice values higher than .50 for both networks even with small mask sizes of 5000 voxels. Good ICC values were also observed for both networks in whole-brain analyses and especially within ROIs. This superiority of within-ROI in relation to whole-brain ICCs is consistent with previous reports (e.g., Pinter et al., 2016), and it is not surprising as the voxels within ROIs are supposed to be directly linked with the networks under study.

Sensorimotor networks were found to be more reliable than language networks using Dice coefficients (at fixed *z* scores, and fixed mask sizes). This goes in line with previous evidence that mapping higher-level cognitive networks is less reliable than mapping sensorimotor and visual networks (Chen et al., 2015; Pinter et al., 2016; Shirer et al., 2015). Larger variability related to subject-specific states during scanning might have played a role in our result of less reliability for the language in comparison with the sensorimotor network. Similar reliability estimates were observed for short (hours) and medium (7 days) timescales in both networks. This is coherent with previous reports that rs-fMRI networks are stable over periods of days to weeks (Mannfolk et al., 2011; Meindl et al., 2010). Other studies have demonstrated reliable rs-fMRI networks within larger timescales of months (5–16 months: Chou et al. (2012); Zhu et al. (2014)) and even years in individuals scanned multiple times over longer periods (1 year: Laumann et al. (2015); 3.5 years: Choe et al. (2015)). So our findings add to converging evidence on the temporal stability of rs-fMRI networks.

How does the magnitude of the reliability estimates found here compare with values found with conventional and better studied task-based fMRI? In a review of task-based test-retest reliability with various tasks, Bennett and Miller (2010) point to average values of 0.45 for Dice and 0.50 for ICC. In a study of single-subject test-retest metrics, Gorgolewski et al. (2013) found average whole-brain Dice values of 0.515 for motor mapping (finger tapping), and of 0.502 and 0.452 for language mapping with verb-generation and verb-repetition, respectively. More recently, Morrison et al. (2016) reported single-subject average whole-brain Dice values of 0.55 for a hand-squeezing task, 0.49 for a phonemic fluency task and 0.56 for a rhyme judgment task. Across these studies, the reliability of task-based fMRI fell between 0.31 to 0.67 when measured with Dice coefficients and between 0.17 to 0.75 when measured with ICC. Our reliability results here were in the upper range of those values, suggesting that ultra-high field rs-fMRI can be as reliable as task-based approaches. Interestingly, in a direct comparison of single-subject task-based vs. rs-fMRI test-retest metrics, Kristo et al. (2014) found an ICC of 0.42 for task-based fMRI but only 0.25 for rs-fMRI. Here, we obtained average ICC values almost three times higher. There are several differences between the rs-fMRI methods of two studies: Kristo et al. (2014) used a seed-based correlation approach whereas we resorted to ICA and dual-regression analysis; their length of acquisition time was 4.05 min (400 volumes) whereas ours was 15 min (300 vol), and it is known that longer scanning times improve reliability (Birn et al., 2013). However, an important difference is that that Kristo et al.'s data were acquired at 3 T and ours at 7 T.

One of the key aspects of the present study is the use of ultra-high resolution. The main requirement for accurate reliability estimates is a precise delineation of the networks under study, which is not warranted if large voxel sizes or large amounts of spatial smoothing are used — an unavoidable constraint with lower magnetic fields. The good reliability estimates that we have observed here with 7 T MRI hold great promise for the clinical applications of ultra-high field resolution imaging. However, one should also be aware of its limitations. These include heightened sensitivity to motion and physiological artefacts, dielectric resonance effects, as well as magnetic susceptibility artefacts that may cause geometric distortions in the functional images; all of these hinder the accuracy of the image registration procedure. Indeed, we found that coregistration error was the only significant confound that affected the reliability metrics measured with a one week interval (in the sensorimotor network). Besides, coregistration errors were larger between sessions than within session: a threefold increase in relation to the two protocols of the first session, where the subject did nor leave the scanner and so head repositioning was not needed, and 1.5 times larger in relation to protocols of the second session, where head repositioning occurred. Furthermore, while in a same-day scanning session just one fieldmap was acquired and used for both

fMRI runs, between sessions a new fieldmap was used. The introduction of a second fieldmap and potential differences in field inhomogeneity might have impaired the quality of coregistration and thus reduce reliability. Interestingly, in a previous study with 3 T imaging (Gorgolewski et al., 2013) between-session coregistration error was not significantly related with reliability metrics. This is consistent with our interpretation that working with 7 T data poses important technical challenges regarding the correction of geometric distortions induced by magnetic field inhomogeneities. It is worth mentioning that the impact of coregistration error was detected only in the sensorimotor network; this network might be more susceptible to coregistration error and field inhomogeneity than the language network because of its more restricted spatial topography and closeness to the cranium.

We did not observe a significant contribution of motion or temporal SNR to the reliability metrics. With a small sample size as here (20 subjects), this result must be interpreted with caution. A plausible interpretation is that it is at least in part due to recent advances in artefact removal including ICA denoising, which substantially improve the quality of fMRI analyses and rs-fMRI reproducibility (Pruim et al., 2015b). Although definite conclusions cannot be drawn from our null result, it is tempting to put forward the hypothesis that motion sensitivity of 7 T MRI is not a critical issue for temporal reliability if adequate pre-processing steps are taken. Similarly, we found no relationship between subject-related variables (sleep habits, hydration, pulse and blood pressure) and reliability metrics. This suggests that these variables contribute little to the reliability of rs-fMRI as measured here. Previous studies have reported changes in rs-fMRI connectivity related to fluctuations over time in psychological, physical and metabolic variables (Laumann et al., 2015; Poldrack et al., 2015). It remains to be seen whether such changes are large enough to impact on single-subject network mapping, and these questions deserve to be explored in the future.

In conclusion, our results provide strong support for the suitability of rs-fMRI for single-subject mapping, adding to the evidence recently provided by Mannfolk et al. (2011), Rosazza et al. (2014), and Tie et al. (2014). We demonstrated for the first time that rs-fMRI at ultra-high resolution could reliably map two important networks, the sensorimotor and the language networks. This sets the ground for future ultra-high field clinical applications that require single-subject mapping including preoperative planning. Whether this reliability can be obtained in clinical samples should be the focus of future studies.

Acknowledgments

This work was supported by grants from the Portuguese Foundation for Science and Technology (SFRH/BD/86912/2012; UID/PSI/000050/2013). We are grateful to K. Gorgolewski and colleagues (2015) for providing the open data. We thank Ricardo Pereira for helping with data processing. We also thank K. Gorgolewski and an anonymous reviewer for valuable contributions to this manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuroimage.2016.11.029.

Supporting material.

In-depth instructions and supporting scripts are provided at github (https://github.com/brancop/Brancoetal_7tesla_testretest/) to allow the replication of the analyses included in this article with minor user-input.

References

Andersson, J.L., Jenkinson, M., Smith, S., 2007. Non-linear Registration, Aka Spatial Normalization, (FMRIB Technical Report TR07JA2). Retrieved from ⟨www.fmrib.ox.ac.uk/analysis/techrep⟩.

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360 (1457), 1001–1013. http://dx.doi.org/10.1098/rstb.2005.1634.

Beckmann, C.F., Mackay, C.E., Filippini, N., Smith, S.M., 2009. Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. Neuroimage 47 (Suppl 1), S148. http://dx.doi.org/10.1016/s1053-8119(09)71511-3.

Beckmann, C.F., Smith, S.M., 2004. Probabilistic independent component analysis for functional magnetic resonance imaging. IEEE Trans. Med. Imaging 23 (2), 137–152. http://dx.doi.org/10.1109/tmi.2003.822821.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191 (1), 133–155. http://dx.doi.org/10.1111/j.1749-6632.2010.05446.x.

Birn, R.M., Molloy, E.K., Patriat, R., Parker, T., Meier, T.B., Kirk, G.R., Prabhakaran, V., 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. NeuroImage 83, 550–558. http://dx.doi.org/10.1016/j.neuroimage.2013.05.099.

Branco, P., Seixas, D., Deprez, S., Kovacs, S., Peeters, R., Castro, S.L., Sunaert, S., 2016. Resting-state functional magnetic resonance imaging for language preoperative planning. Front. Hum. Neurosci., 10. http://dx.doi.org/10.3389/fnhum.2016.00011.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45 (3), 758–768. http://dx.doi.org/10.1016/j.neuroimage.2008.12.035.

Camchong, J., MacDonald, A.W., Bell, C., Mueller, B.A., Lim, K.O., 2009. Altered functional and anatomical connectivity in schizophrenia. Schizophr. Bull. 37 (3), 640–650. http://dx.doi.org/10.1093/schbul/sbp131.

Chen, B., Xu, T., Zhou, C., Wang, L., Yang, N., Wang, Z., Weng, X.C., 2015. Individual variability and test-retest reliability revealed by ten repeated resting-state brain scans over one month. PLoS One 10 (12), e0144963. http://dx.doi.org/10.1371/journal.pone.0144963.

Choe, A.S., Jones, C.K., Joel, S.E., Muschelli, J., Belegu, V., Caffo, B.S., Pekar, J.J., 2015. Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years. PLoS One 10 (10), e0140134. http://dx.doi.org/10.1371/journal.pone.0140134.

Chou, Y.H., Panych, L.P., Dickey, C.C., Petrella, J.R., Chen, N.K., 2012. Investigation of long-term reproducibility of intrinsic connectivity network mapping: a resting-state fMRI study. Am. J. Neuroradiol. 33 (5), 833–838. http://dx.doi.org/10.3174/ajnr.a2894.

Cochereau, J., Deverdun, J., Herbet, G., Charroud, C., Boyer, A., Moritz-Gasser, S., Duffau, H., 2016. Comparison between resting state fMRI networks and responsive cortical stimulations in glioma patients. Hum. Brain Mapp.. http://dx.doi.org/10.1002/hbm.23270, (Advance online publication).

De Martino, F., Esposito, F., van de Moortele, P.F., Harel, N., Formisano, E., Goebel, R., Yacoub, E., 2011. Whole brain high-resolution functional imaging at ultra high magnetic fields: an application to the analysis of resting state networks. NeuroImage 57 (3), 1031–1044. http://dx.doi.org/10.1016/j.neuroimage.2011.05.008.

DeSalvo, M.N., Tanaka, N., Douw, L., Leveroni, C.L., Buchbinder, B.R., Greve, D.N., Stufflebeam, S.M., 2016. Resting-state functional MR imaging for determining language laterality in intractable epilepsy. Radiology, 141010. http://dx.doi.org/10.1148/radiol.2016141010.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Albert, M.S., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31 (3), 968–980. http://dx.doi.org/10.1016/j.neuroimage.2006.01.021.

Fedorenko, E., Hsieh, P.J., Nieto-Castañón, A., Whitfield-Gabrieli, S., Kanwisher, N., 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104 (2), 1177–1194. http://dx.doi.org/10.1152/jn.00032.2010.

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60 (6), 969–975. http://dx.doi.org/10.1212/01.wnl.0000049934.34209.2e.

Filippini, N., MacIntosh, B.J., Hough, M.G., Goodwin, G.M., Frisoni, G.B., Smith, S.M., Mackay, C.E., 2009. Distinct patterns of brain activity in young carriers of the APOE-ε4 allele. Proc. Natl. Acad. Sci. USA 106 (17), 7209–7214. http://dx.doi.org/10.1016/s1053-8119(09)71381-3.

Fox, M.D., Qian, T., Madsen, J.R., Wang, D., Li, M., Ge, M., Liu, H., 2016. Combining task-evoked and spontaneous activity to improve pre-operative brain mapping with fMRI. NeuroImage 124, 714–723. http://dx.doi.org/10.1016/j.neuroimage.2015.09.030.

Gorgolewski, K.J., Mendes, N., Wilfling, D., Wladimirow, E., Gauthier, C.J., Bonnen, T., Smallwood, J., 2015. A high resolution 7 T resting-state fMRI test-retest dataset with cognitive and physiological measures. Sci. Data 2, 140054. http://dx.doi.org/10.1038/sdata.2014.54.

Gorgolewski, K., Storkey, A.J., Bastin, M.E., Pernet, C.R., 2012. Adaptive thresholding for reliable topological inference in single subject fMRI analysis. Front. Hum. Neurosci. 6, 245. http://dx.doi.org/10.3389/fnhum.2012.00245.

Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013. Single subject fMRI test–retest reliability metrics and confounding factors. NeuroImage 69,

231–243. http://dx.doi.org/10.1016/j.neuroimage.2012.10.085.
Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48 (1), 63–72. http://dx.doi.org/10.1016/j.neuroimage.2009.06.060.
Grömping, U., 2006. Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. 17 (1), 1–27. http://dx.doi.org/10.18637/jss.v017.i01.
Hale, J.R., Brookes, M.J., Hall, E.L., Zumer, J.M., Stevenson, C.M., Francis, S.T., Morris, P.G., 2010. Comparison of functional connectivity in default mode and sensorimotor networks at 3 and 7 T. Magn. Reson. Mater. Phys. Biol. Med. 23 (5-6), 339–349. http://dx.doi.org/10.1007/s10334-010-0220-0.
Hall, W.A., Liu, H., Truwit, C.L., 2005. Functional magnetic resonance imaging–guided resection of low-grade gliomas. Surg. Neurol. 64 (1), 20–27. http://dx.doi.org/10.1016/j.surneu.2004.08.099.
Jenkinson, M., 2003. Fast, automated, N-dimensional phase-unwrapping algorithm. Magn. Reson. Med. 49 (1), 193–197. http://dx.doi.org/10.1002/mrm.10354.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841. http://dx.doi.org/10.1006/nimg.2002.1132.
Kristo, G., Rutten, G.J., Raemaekers, M., Gelder, B., Rombouts, S.A., Ramsey, N.F., 2014. Task and task-free FMRI reproducibility comparison for motor network identification. Hum. Brain Mapp. *35* (1), 340–352. http://dx.doi.org/10.1002/hbm.22180.
Kundu, B., Penwarden, A., Wood, J.M., Gallagher, T.A., Andreoli, M.J., Voss, J., Moritz, C., 2013. Association of functional magnetic resonance imaging indices with postoperative language outcomes in patients with primary brain tumors. Neurosurg. Focus 34 (4), E6. http://dx.doi.org/10.3171/2013.2.focus12413.
Laumann, T.O., Gordon, E.M., Adeyemo, B., Snyder, A.Z., Joo, S.J., Chen, M.Y., Schlaggar, B.L., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87 (3), 657–670. http://dx.doi.org/10.1016/j.neuron.2015.06.037.
Mahowald, K., Fedorenko, E., 2016. Reliable individual-level neural markers of high-level language processing: a necessary precursor for relating neural variability to behavioral and genetic variability. NeuroImage. http://dx.doi.org/10.1016/j.neuroimage.2016.05.073, (Advance online publication).
Mannfolk, P., Nilsson, M., Hansson, H., Ståhlberg, F., Fransson, P., Weibull, A., Olsrud, J., 2011. Can resting-state functional MRI serve as a complement to task-based mapping of sensorimotor function? A test–retest reliability study in healthy volunteers. J. Magn. Reson. Imaging 34 (3), 511–517. http://dx.doi.org/10.1002/jmri.22654.
Meindl, T., Teipel, S., Elmouden, R., Mueller, S., Koch, W., Dietrich, O., Glaser, C., 2010. Test–retest reproducibility of the default-mode network in healthy individuals. Hum. Brain Mapp. 31 (2), 237–246. http://dx.doi.org/10.1002/hbm.20860.
Mitchell, T.J., Hacker, C.D., Breshears, J.D., Szrama, N.P., Sharma, M., Bundy, D.T., Leuthardt, E.C., 2013. A novel data-driven approach to preoperative mapping of functional cortex using resting-state functional magnetic resonance imaging. Neurosurgery 73 (6), 969. http://dx.doi.org/10.1227/neu.0000000000000141.
Morrison, M.A., Churchill, N.W., Cusimano, M.D., Schweizer, T.A., Das, S., Graham, S.J., 2016. Reliability of task-based fMRI for preoperative planning: a test-retest study in brain tumor patients and healthy controls. PLoS One 11 (2), e0149547. http://dx.doi.org/10.1371/journal.pone.0149547.
Petrella, J.R., Shah, L.M., Harris, K.M., Friedman, A.H., George, T.M., Sampson, J.H., Voyvodic, J.T., 2006. Preoperative functional MR imaging localization of language and motor areas: effect on therapeutic decision making in patients with potentially resectable brain tumors. Radiology 240 (3), 793–802. http://dx.doi.org/10.1148/radiol.2403051153.
Pinter, D., Beckmann, C., Koini, M., Pirker, E., Filippini, N., Pichler, A., Enzinger, C., 2016. Reproducibility of resting state connectivity in patients with stable multiple sclerosis. PLoS One 11 (3), e0152158. http://dx.doi.org/10.1371/journal.pone.0152158.
Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.Y., Hunicke-Smith, S., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun., 6. http://dx.doi.org/10.1038/ncomms9885.
Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage 84, 320–341. http://dx.doi.org/10.1016/j.neuroimage.2013.08.048.
Price, C.J., Crinion, J., Friston, K.J., 2006. Design and analysis of fMRI studies with neurologically impaired patients. J. Magn. Reson. Imaging 23 (6), 816–826. http://dx.doi.org/10.1002/jmri.20580.
Pruim, R.H., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015a. ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage 112, 267–277. http://dx.doi.org/10.1016/j.neuroimage.2015.02.064.
Pruim, R.H., Mennes, M., Buitelaar, J.K., Beckmann, C.F., 2015b. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. NeuroImage 112, 278–287. http://dx.doi.org/10.1016/j.neuroimage.2015.02.063.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F., 2007. Test-retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542. http://dx.doi.org/10.1016/j.neuroimage.2007.03.061.
Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imaging 16 (2), 105–113. http://dx.doi.org/10.1016/s0730-725x(97)00253-1.
Rosazza, C., Aquino, D., D'Incerti, L., Cordella, R., Andronache, A., Zacà, D., Minati, L., 2014. Preoperative mapping of the sensorimotor cortex: comparative assessment of task-based and resting-state FMRI. PLoS One 9 (6), e98860. http://dx.doi.org/10.1371/journal.pone.0098860.
Sair, H.I., Yahyavi-Firouz-Abadi, N., Calhoun, V.D., Airan, R.D., Agarwal, S., Intrapiromkul, J., Pillai, J.J., 2016. Preoperative brain mapping of the language network in patients with brain tumors using resting-state fMRI: comparison with task fMRI. Hum. Brain Mapp. 37 (3), 913–923. http://dx.doi.org/10.1002/hbm.23075.
Sanchez-Panchuelo, R.M., Francis, S., Bowtell, R., Schluppeck, D., 2010. Mapping human somatosensory cortex in individual subjects with 7 T functional MRI. J. Neurophysiol. 103 (5), 2544–2556. http://dx.doi.org/10.1152/jn.01017.2009.
Shehzad, Z., Kelly, A.C., Reiss, P.T., Gee, D.G., Gotimer, K., Uddin, L.Q., Petkova, E., 2009. The resting brain: unconstrained yet reliable. Cereb. Cortex 19 (10), 2209–2229. http://dx.doi.org/10.1093/cercor/bhn256.
Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D., 2012. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb. Cortex 22 (1), 158–165. http://dx.doi.org/10.1093/cercor/bhr099.
Shirer, W.R., Jiang, H., Price, C.M., Ng, B., Greicius, M.D., 2015. Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. NeuroImage 117, 67–79. http://dx.doi.org/10.1016/j.neuroimage.2015.05.015.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155. http://dx.doi.org/10.1002/hbm.10062.
Sunaert, S., 2006. Presurgical planning for tumor resectioning. J. Magn. Reson. Imaging 23 (6), 887–905. http://dx.doi.org/10.1002/jmri.20582.
Tie, Y., Rigolo, L., Norton, I.H., Huang, R.Y., Wu, W., Orringer, D., Golby, A.J., 2014. Defining language networks from resting-state fMRI for surgical planning—a feasibility study. Hum. Brain Mapp. 35 (3), 1018–1030. http://dx.doi.org/10.1002/hbm.22231.
Wengenroth, M., Blatow, M., Guenther, J., Akbar, M., Tronnier, V.M., Stippich, C., 2011. Diagnostic benefits of presurgical fMRI in patients with brain tumours in the primary sensorimotor cortex. Eur. Radiol. 21 (7), 1517–1525. http://dx.doi.org/10.1007/s00330-011-2067-9.
Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., Nichols, T.E., 2014. Permutation inference for the general linear model. NeuroImage 92, 381–397. http://dx.doi.org/10.1016/j.neuroimage.2014.01.060.
Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Smith, S.M., 2005. Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. IEEE Trans. Med. Imaging 24 (1), 1–11. http://dx.doi.org/10.1109/tmi.2004.836545.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57. http://dx.doi.org/10.1109/42.906424.
Zhu, L., Fan, Y., Zou, Q., Wang, J., Gao, J.H., Niu, Z., 2014. Temporal reliability and lateralization of the resting-state language network. PLoS One 9 (1), e85880. http://dx.doi.org/10.1371/journal.pone.0085880.
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. http://dx.doi.org/10.1016/j.neubiorev.2014.05.009.
Zuo, X.N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Chen, A., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 1, 140049. http://dx.doi.org/10.1038/sdata.2014.49.