NeuroImage 169 (2018) 172–175

Contents lists available at ScienceDirect

NeuroImage

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

Considering factors affecting the connectome-based identification process: Comment on Waller et al.

Corey Horien [a,*], Stephanie Noble [a], Emily S. Finn [a,d], Xilin Shen [b], Dustin Scheinost [b,c], R. Todd Constable [a,b,e]

[a] Interdepartmental Neuroscience Program, Yale University School of Medicine, New Haven, CT, USA
[b] Department of Radiology and Biomedical Imaging, Yale University School of Medicine, New Haven, CT, USA
[c] The Child Study Center, Yale University School of Medicine, New Haven, CT, USA
[d] Section on Functional Imaging Methods, National Institute of Mental Health, Bethesda, MD, USA
[e] Department of Neurosurgery, Yale University School of Medicine, New Haven, CT, USA

A B S T R A C T

A recent study by Waller and colleagues evaluated the reliability, specificity, and generalizability of using functional connectivity data to identify individuals from a group. The authors note they were able to replicate identification rates in a larger version of the original Human Connectome Project (HCP) dataset. However, they also report lower identification accuracies when using historical neuroimaging acquisitions with low spatial and temporal resolution. The authors suggest that their results indicate connectomes derived from historical imaging data may be similar across individuals, to the extent that this connectome-based approach may be inappropriate for precision psychiatry and the goal of drawing inferences based on subject-level data. Here we note that the authors did not take into account factors affecting data quality and hence identification rates, independent of whether a low spatiotemporal resolution acquisition or a high spatiotemporal resolution acquisition is used. Specifically, we show here that the amount of data collected per subject and in-scanner motion are the predominant factors influencing identification rates, not the spatiotemporal resolution of the acquisition. To do this, we investigated identification rates in the HCP dataset as a function of the amount of data and motion. Using a dataset from the Consortium for Reliability and Reproducibility (CoRR), we investigated the impact of multiband versus non-multiband imaging parameters; that is, high spatiotemporal resolution versus low spatiotemporal resolution acquisitions. We show scan length and motion affect identification, whereas the imaging protocol does not affect these rates. Our results suggest that motion and amount of data per subject are the primary factors impacting individual connectivity profiles, but that within these constraints, individual differences in the connectome are readily observable.

Introduction

A key goal of precision psychiatry is leveraging individual differences in neuroimaging data to generate predictive models related to behavior. As highlighted by Waller et al. (2017), finding reliable markers across datasets remains an important part of this process. As such, they investigate the generalizability of a previous method using functional connectivity fMRI data to identify individuals from a group ('connectome fingerprinting'; Finn et al., 2015). Waller et al. demonstrate that identification can be replicated in the same high spatiotemporal resolution dataset (i.e. acquired using multiband acquisition sequences), consistent with other work to replicate the method (Finn et al., 2017; Kaufmann et al., 2017; Vanderwal et al., 2017), though they note lower accuracies using a dataset acquired with lower spatiotemporal resolution (i.e. acquired using non-multiband acquisition sequences). In addition, the authors also show that the specificity of the identification procedure is lower when a within-subject correlation threshold is introduced into the ID pipeline. Therefore, the authors argue that the identification method may not generalize to datasets with lower spatiotemporal resolution because individual features may only be detectable in data acquired with high spatiotemporal resolution. However, in their study, the authors did not take into account other factors affecting data quality and hence the identification process, namely scan duration and subject motion. Here we evaluate the impact of not only spatiotemporal resolution during image acquisition, but also other data quality factors on identification rates.

Methods

The HCP 900 subjects release (Van Essen et al., 2013) was used to investigate scan time and motion. Data were pre-processed and connectivity matrices were calculated as described elsewhere (Finn et al., 2015, 2017; Shen et al., 2017). All analyses were performed using the left-to-right (LR) phase encoding rest runs from days one and two. Of note, HCP TR = 720 ms. To study motion, subjects were separated into

* Corresponding author. Magnetic Resonance Research Center, 300 Cedar St, PO Box 208043, New Haven, CT 06520-8043, USA.
E-mail address: corey.horien@yale.edu (C. Horien).

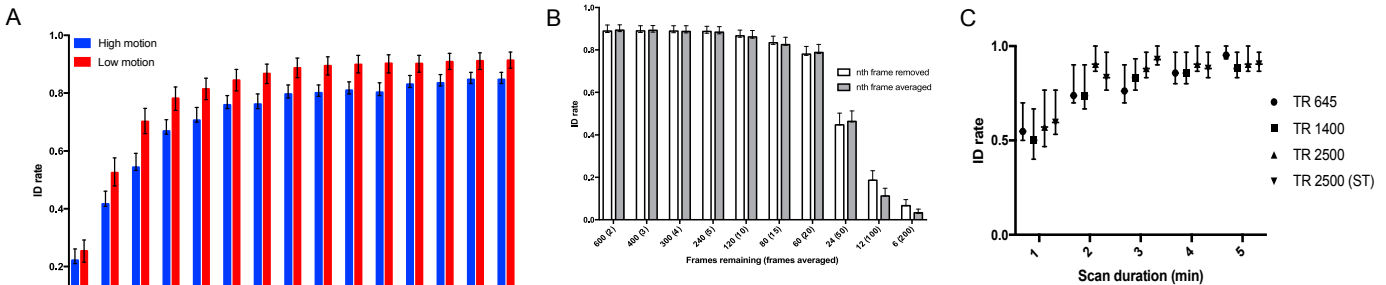
https://doi.org/10.1016/j.neuroimage.2017.12.045
Received 26 September 2017; Accepted 14 December 2017
Available online 15 December 2017

low and high motion groups using a mean frame-frame displacement threshold of 0.1 mm averaged over the two sessions. Of the 819 subjects available with all data and day one and two LR rest scans, 603 were in the low motion group and 216 subjects were in the high motion group. To study the effect of scan time, we truncated time courses to correspond to the number of frames in 1, 2, …,14 min and calculated connectivity matrices from the shortened data. Because of the difference in sample sizes among the low and high motion groups, we repeatedly subsampled 216 subjects in the low motion group and performed identification 1000 times. The mean ID rate and 95% confidence intervals were therefore calculated from the subsampled data (Fig. 1A). To investigate when ID rates plateaued, we used the Levenberg-Marquardt nonlinear least squares algorithm to fit the following nonlinear regression model function: $IDrate = maxIDrate\left(1 - e^{-\frac{t}{x}}\right)$, where t = time, $IDrate$ = ID rate at time t, $maxIDrate$ = maximum ID rate determined by the model, and x = time required for the ID rate to reach approximately 63% of its maximum value. We defined plateauing of the ID rate to be the time points when the rate was 95 and 99% of the maximum ID rate.

In a separate analysis (Fig. 1B), we subsampled data (after low-pass filtering; approximate cutoff frequency of 0.12 Hz) from each of the 603 low motions subjects to simulate the effects of lower sampling frequencies (longer TR) versus total amount of scan time. For this analysis, we selected n frames from the duration of a subject's time course such that sampling every other frame produced 600 frames of the original 1200; sampling every 3rd frame resulted in 400 frames remaining, etc. Hence, these subsampled data still spanned the same overall acquisition time window. It should also be noted that this subsampled data has lower signal to noise ratio (SNR) than real data acquired at a longer TR because of the additional T1 recovery that would occur with a longer TR. Connectivity matrices were subsequently calculated from the subsampled data. In addition, we performed a follow-up analysis using a similar strategy except that instead of removing every nth frame we averaged data from every n adjacent frames to again simulate a slower sampling frequency (Fig. 1B) and boost the SNR.

To study the effect of spatiotemporal resolution (Fig. 1C), we utilized a publically available test-retest dataset from the Nathan Kline Institute (NKI; http://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_1.html). This dataset contains individuals scanned with both multiband and non-multiband acquisition sequences, thus allowing us to investigate the impact of different pulse sequences on ID rates. Data acquisition parameters have been described previously (Liao et al., 2013). Briefly, three resting-state fMRI sequences were obtained for each of the 24 subjects: 1) multiband scan with TR = 645 ms; 2) multiband scan with TR = 1400 ms; and 3) and a non-multiband echo planar imaging (EPI) scan with TR = 2500 ms. One subject was excluded due to brain atrophy (subject 0021001); one subject was excluded due to excessive head motion (3795193; greater than 3° rotation); and we were unable to locate session 2 data for subject 6471972, leaving 21 subjects in the final analysis. We did not apply a further motion cutoff with these subjects due to the small sample size. The preprocessing steps have been previously described (Noble et al., 2017), except we performed skull-stripping using optiBET (Lutkenhoff et al., 2014). Though we did not perform slice-time correction on the multi-band data, we performed analyses on the TR = 2500 subjects with and without slice-time correction.

The identification procedure was carried out as described previously using Matlab code released by Finn et al. (2015) and utilized by Waller



HCP low motion		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.254	0.215	0.292
2	0.525	0.479	0.576
3	0.702	0.660	0.748
4	0.782	0.741	0.822
5	0.815	0.778	0.852
6	0.845	0.808	0.882
7	0.868	0.833	0.900
8	0.887	0.854	0.921
9	0.894	0.863	0.926
10	0.899	0.870	0.931
11	0.903	0.870	0.933
12	0.903	0.873	0.931
13	0.908	0.882	0.938
14	0.912	0.882	0.940
Full data	0.914	0.887	0.942

HCP high motion		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.222	0.211	0.261
2	0.417	0.408	0.461
3	0.544	0.533	0.592
4	0.669	0.658	0.708
5	0.706	0.700	0.750
6	0.759	0.747	0.792
7	0.762	0.747	0.797
8	0.796	0.783	0.828
9	0.801	0.789	0.828
10	0.810	0.797	0.839
11	0.803	0.792	0.836
12	0.831	0.819	0.861
13	0.836	0.825	0.864
14	0.847	0.833	0.872
Full data	0.847	0.833	0.872

HCP low motion		95% confidence interval	
Frames remaining	ID rate	Lower bound	Upper bound
600	0.889	0.883	0.918
400	0.890	0.883	0.914
300	0.889	0.881	0.914
240	0.887	0.879	0.911
120	0.867	0.860	0.894
80	0.834	0.826	0.865
60	0.780	0.774	0.818
24	0.448	0.451	0.503
12	0.187	0.190	0.231
6	0.066	0.065	0.095

HCP low motion		95% confidence interval	
Frames averaged	ID rate	Lower bound	Upper bound
2	0.894	0.886	0.919
3	0.892	0.884	0.915
4	0.887	0.880	0.914
5	0.884	0.876	0.910
10	0.862	0.855	0.891
15	0.825	0.821	0.860
20	0.789	0.784	0.826
50	0.464	0.461	0.513
100	0.112	0.114	0.149
200	0.033	0.030	0.050

TR 645		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.548	0.500	0.700
2	0.738	0.700	0.900
3	0.762	0.700	0.900
4	0.857	0.800	0.967
5	0.952	0.933	1.000

TR 1400		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.500	0.400	0.667
2	0.738	0.667	0.900
3	0.833	0.767	0.933
4	0.857	0.800	0.967
5	0.881	0.833	0.967

TR 2500		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.571	0.467	0.767
2	0.905	0.867	1.000
3	0.881	0.833	0.967
4	0.905	0.867	1.000
5	0.905	0.867	1.000

TR 2500 (ST)		95% confidence interval	
Time (min)	ID rate	Lower bound	Upper bound
1	0.595	0.533	0.767
2	0.833	0.767	0.967
3	0.929	0.900	1.000
4	0.881	0.833	0.967
5	0.905	0.867	0.967

Fig. 1. The effect of scan duration, motion, and differences in spatiotemporal resolution on identification rates. (A) Top: Separating the HCP 900 subjects into groups based on motion and performing identification with increasing amounts of data. Identification rate for each group is indicated at each scan duration time. Note that the high and low motion groups have the same scan durations at a given time point on the x-axis. Both groups have equal sample sizes (n = 216). Red and blue bars represent the low and high motion groups, respectively. (B) Top: Simulating the effects of a lower TR in the HCP. Data from all 603 low motion subjects were subsampled (white bars); x-axis indicates the number of frames remaining. In a separate analysis every n adjacent frames were averaged (grey bars); x-axis indicates in parentheses the number of adjacent frames used to average. (C) Top: Identification rates achieved using multiband or non-multiband imaging parameters to assess the effect of spatiotemporal resolution. Multiband imaging was performed on groups labelled as TR 645 and TR 1400; TR 2500 was acquired via non-multiband imaging; TR 2500 (ST) indicates these subjects underwent slice-time correction. Identification rate achieved for each scanning protocol is indicated at each scan duration time. Error bars correspond to 95% confidence intervals. Note that in (A), (B), and (C), the lower part of each panel includes the actual ID rate obtained and the 95% confidence intervals.