An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI.

Linden Parkes[1*], Ben Fulcher[1], Murat Yücel[1], Alex Fornito[1]

[1]Brain & Mental Health Laboratory,
Monash Institute of Cognitive and Clinical Neurosciences and School of Psychological Sciences, Monash University, Victoria, Australia

*Corresponding author: linden.parkes@monash.edu, 770 Blackburn Road, Clayton, Victoria 3168, Australia

Abbreviated title: Motion correction in resting-state fMRI

Author contributions: L. P., B. F., and A. F. designed research; L. P. performed research; L. P., B. F., and A. F. contributed unpublished reagents/analytic tools; L. P., B. F., and A. F. analyzed data; L. P., B. F., M. Y., and A. F. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

Keywords: fMRI, functional connectivity, resting-state, motion, noise, artefact

Acknowledgements: L.P. was supported by an Australian Postgraduate Award and the David Winston Turner Endowment Fund. B.F. was supported by a National Health and Medical Research Council Early Career Fellowship (ID: 1089718). M.Y. was supported by a National Health and Medical Research Council Fellowship (ID: 1117188), Monash University and the David Winston Turner Endowment Fund. A.F. was supported by an Australian Research Council Future Fellowship (ID: FT130100589) and National Health and Medical Research Council Project grants (ID: 3251213, 3251250, 3251392).

Abstract

Estimates of functional connectivity derived from resting-state functional magnetic resonance imaging (rs-fMRI) are highly sensitive to artefacts caused by in-scanner head motion. This susceptibility has motivated the development of numerous denoising methods designed to mitigate motion-related artefacts. Here, we compare 8 popular retrospective rs-fMRI denoising methods, including methods such as regression of head motion parameters (with or without expansion terms), aCompCor, volume censoring (e.g., scrubbing and spike regression), global signal regression and ICA-AROMA, combined into 16 different pipelines. These pipelines were evaluated across five different quality control benchmarks in three independent datasets that were characterized by both high and low levels of motion. Pipelines were benchmarked by examining the residual relationship between in-scanner movement and functional connectivity after denoising; the effect of distance on this residual relationship; whole-brain differences in functional connectivity between high- and low-motion healthy controls (HC); the temporal degrees of freedom lost during denoising; and the test-retest reliability of functional connectivity estimates. We also compared the sensitivity of each pipeline to clinical differences in functional connectivity in comparisons between people with schizophrenia (SCZ; n = 50) and HCs (n = 121) and people with obsessive-compulsive disorder (OCD; n = 34) and HCs (n = 39). Our results indicate that (1) simple linear regression of regional fMRI time series against head motion parameters (with or without expansion terms) is not sufficient to remove head motion artefacts; (2) aCompCor pipelines can exacerbate motion artefacts in low-motion data; (3) the primary benefit of volume censoring comes from the exclusion of high-motion individuals rather than censoring of data in remaining participants; and (4) that ICA-AROMA consistently performed well across all benchmarks and datasets, particularly when applied after the exclusion of high-motion individuals. ICA-AROMA was also the most sensitive to clinical differences in case-control analyses, suggesting that its denoising efficacy is associated with enhanced power for detecting pathophysiological effects. Crucially, the comparison between HC and SCZ revealed that the specific choice of noise correction pipeline had a major effect on the findings, affecting both the location and direction of group differences. Putative increases in functional connectivity in patients only emerged in pipelines incorporating either global signal regression or aCompCor. Thus, group comparisons in functional connectivity are highly dependent on preprocessing strategy. We offer some recommendations for best practice and outline some simple analyses to facilitate transparent reporting of the degree to which a given set of findings may be affected by motion-related artefact.

Introduction

Fluctuations of the blood-oxygenation-level-dependent (BOLD) signal recorded with functional magnetic resonance imaging during task-free "resting state" experiments (rs-fMRI) are highly organized, being correlated across anatomically distributed networks (Fox and Raichle, 2007) that correspond to those typically co-activated during task performance (Smith et al., 2009). These spontaneous dynamics predict task-evoked activation and behaviour (Cole et al., 2014; Fox and Raichle, 2007; Fox et al., 2007), can be used to accurately identify individuals across repeated scans (Finn et al., 2015), and are under significant genetic control (Fornito et al., 2011b; Glahn et al., 2010). These findings suggest that resting-state fMRI can be used to probe a functionally important aspect of intrinsic brain dynamics which, together with the relative ease of data acquisition, has made the technique an attractive phenotyping tool for studies of brain disease and at-risk populations (Dandash et al., 2014; Fornito and Bullmore, 2010; Fornito et al., 2013).

A major obstacle in the analysis of fMRI data, particularly those acquired during unconstrained resting-state conditions, is contamination of the BOLD signal by head motion and fluctuations in non-neuronal physiological processes. Head motion is a particularly pernicious problem. Even small movements of the head between volumes acquired during a scan will cause a voxel to sample different brain regions over time, thus compromising accurate measurement of brain dynamics. This contamination can influence estimates of functional connectivity – i.e., statistical estimates of pairwise time series covariation – such that increased motion characteristically attenuates long-range coupling and inflates short-range coupling between brain regions (Power et al., 2012; Van Dijk et al., 2012). The problem is especially problematic in group comparisons (e.g., between patients and controls), where differences in head motion can introduce systematic bias in connectivity estimates.

The most commonly-used method for removing motion-related noise from recorded BOLD signals is linear regression (Fox and Raichle, 2007). With this approach, voxel-wise BOLD time series are regressed against head motion time series estimated along six dimensions (i.e., translational displacements along the *X*-, *Y*-, and *Z*-axes, and rotational displacements of pitch, roll, and yaw; hereafter referred to as head motion parameters, or HMP). To control for fluctuations in non-neuronal physiology, it is also common to regress voxel-averaged time courses extracted from tissue compartments thought to contain nuisance signals, such as white matter (WM) and cerebrospinal fluid (CSF) (Fox and Raichle, 2007). The residuals of this confound regression are then used for further analysis. Expansion terms

can also be added to the model to account for residual variance not removed by first-order effects. For example, Friston et al. (1996) recommended an autoregressive model that also included the motion estimates from the previous time point, as well as square terms (Friston-24 model; M_t, M_{t-1}, M_t^2, M_{t-1}^2, where M is the motion time series for a given dimension and t is time). Other researchers have used expanded models that incorporate temporal derivatives, calculated as backwards differences (Van Dijk et al., 2012), or have included both temporal derivatives and square terms (Satterthwaite et al., 2013).

Another common yet controversial method for correcting for physiological noise and head motion is global signal regression (GSR). GSR corrects for covariance between voxel-wise BOLD signals and the mean BOLD signal averaged across all voxels. GSR has been shown to reduce non-neuronal sources of physiological variance in the BOLD signal, such as those linked to respiration (Birn, 2012), and to mitigate the effects of in-scanner movement (Power et al., 2014; Yan et al., 2013a). However, GSR may also remove BOLD signal fluctuations of neuronal origin (Chen et al., 2012), spuriously weakening some correlations. The method also changes the distribution of functional connectivity estimates in the brain so that it is approximately centred on zero, which causes the emergence of negative correlations (Fox et al., 2009), and which can drive artefactual group differences in functional connectivity (Gotts et al., 2013; Saad et al., 2012). Critically, the extent to which GSR removes noise or signal from BOLD data may be contingent on the amount of global noise present (Chen et al., 2012), which further complicates group comparisons since group differences in head motion (and potentially, non-neuronal physiology) will cause differences in noise levels and thus lead to GSR exerting a differential effect on BOLD time series.

Using the various covariates described above in a regression model can reduce noise in BOLD data, but the overall effects of subtle in-scanner movements from volume-to-volume, called framewise displacements (FDs), are not fully removed by this approach (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). One solution, proposed by Power and colleagues (Power et al., 2014; 2012; 2013; 2015), is called “scrubbing”, and involves removing data points (volumes) for which the FD exceeds a pre-determined threshold. Adjacent volumes are also optionally removed, since large FDs can affect the BOLD signal in preceding and subsequent time points. A related strategy is called spike regression (Lemieux et al., 2007; Satterthwaite et al., 2013), which involves modelling the influence of contaminated time points using separate delta functions, one for each contaminated time point, and removing these effects via linear regression. Several groups have shown that these methods can effectively mitigate the impact of FDs on measures of functional connectivity estimates (Power

et al., 2012; Satterthwaite et al., 2013; Yan et al., 2013a), yet they come at the cost of potentially large amounts of lost data.

To get around the limitations of volume censoring, several alternative, data-driven methods have emerged that can reduce noise without data censoring (Behzadi et al., 2007; Muschelli et al., 2014; Pruim et al., 2015b). One popular method is CompCor (Behzadi et al., 2007), in which BOLD time series from voxels presumed to sample non-neuronal physiology are summarised as temporal principal components and entered as nuisance parameters in a linear regression model. A popular application of this technique is known as anatomical CompCor (aCompCor; Muschelli et al., 2014), which uses noise-related principal components estimated from WM and CSF voxels time courses. Other recent approaches use automated selection of noise-related components from a spatial independent component analysis (ICA) of the data. For example, ICA-FIX selects noise components based on matches to a manually curated training set of noise components (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). More recently, a simpler method called ICA-AROMA, which does not rely on the establishment of a study-specific training dataset, has been developed for the automatic detection of motion-related components according to a specific set of *a priori* criteria (Pruim et al., 2015b; 2015a). Alternative, methods for noise correction that rely on the frequency content of the BOLD signal itself have also been proposed (Patel et al., 2014).

Several recent evaluation studies have been performed to compare the relative performance of many of these BOLD denoising methods on one or more of several quality-control benchmarks. These benchmarks, and the acronyms used to identify them, are summarised in Table 1. Evaluations of different pipelines using various combinations of these benchmarks have generally failed to converge on a single approach that performs best across all outcome metrics. In general, it appears that, compared to simple first-order linear regression models (that include the 6 HMP as well as mean WM/CSF signals), models that include expansion terms perform better on HLM contrasts and motion-BOLD contrasts, but show reduced TRT (Satterthwaite et al., 2012; Yan et al., 2013a). Adding volume censoring to these expansion models yields further improvements to QC-FC correlations, QC-FC distance-dependence, HLM contrasts, and motion-BOLD contrasts (Ciric et al., 2017; Power et al., 2012; Satterthwaite et al., 2013; Yan et al., 2013a). Moreover, incorporating GSR leads to substantial improvements in motion-BOLD contrasts and QC-FC correlations (Ciric et al., 2017; Yan et al., 2013a), but exaggerates QC-FC distance-dependence (Ciric et al., 2017). One study found that expanded motion regression with aCompCor was effective at reducing FD-DVARS without GSR or scrubbing (Muschelli et al., 2014). Another study found that ICA-

AROMA performed equivalently to expanded motion regression with scrubbing and outperformed expanded motion regression with aCompCor on HLM contrasts, all while yielding reduced tDOF-loss (Pruim et al., 2015a). A subsequent comparison found that expanded regression models plus scrubbing outperformed both ICA-AROMA and aCompCor on QC-FC correlations, but that ICA-AROMA was the only method to show virtually no QC-FC distance-dependence (Ciric et al., 2017).

Table 1. Summary of quality control metrics

Quality control benchmark	**Summary**	**References**
QC-FC correlations	The cross-subject correlation between framewise displacement (FD) and functional connectivity at each pair of regions after noise correction.	(Ciric et al., 2017; Power et al., 2015; 2012; Satterthwaite et al., 2013; 2012)
QC-FC distance dependence	The dependence of QC-FC correlations on the distance between brain regions, given evidence that QC-FC values are higher for regional pairs separated by short distances	(Ciric et al., 2017; Power et al., 2012; Satterthwaite et al., 2012)
Motion-BOLD contrasts	Statistical parametric mapping of the association between FD and voxelwise BOLD time courses, to identify regions showing significant motion contamination.	(Yan et al., 2013a)
high-motion vs low-motion contrasts (HLM contrasts)	The mean difference in functional connectivity between healthy control participants split into high- and low-motion subgroups	(Pruim et al., 2015a; Satterthwaite et al., 2013)
FD-DVARS correlations	The cross-subject correlation between motion and the temporal Derivative of root mean square VARiance over voxelS (DVARS), which indexes the rate of change of BOLD signal across the entire brain between consecutive time points.	(Muschelli et al., 2014)
tDOF-loss	The loss in temporal degrees of freedom (tDOF) sustained due to noise correction, calculated as the number of nuisance regressors input to the GLM used to model noise in the BOLD data.	(Ciric et al., 2017; Yan et al., 2013a)
Test-retest reliability (TRT)	The test-retest reliability of functional connectivity, quantified using intra-class correlation coefficients in longitudinally acquired data.	(Birn et al., 2014; Van Dijk et al., 2012; Yan et al., 2013a)

Together, these results suggest that it is difficult to find a single denoising method that performs well across all quality control measures, and that there is a general trade-off between adequately modelling the contributions of noise to the data and limiting the number of noise regressors to avoid over-fitting and/or severe tDOF-loss. However, variability across studies

in terms of the benchmark measures used makes it difficult to perform fair comparisons. Furthermore, none of the benchmarking studies reported thus far have examined how these noise correction strategies impact the analysis of functional connectivity differences between groups, which is a key application of rs-fMRI.

In this study, we evaluated 8 popular rs-fMRI denoising strategies, combined into 16 different pipelines, applied to three independent datasets with respect to five benchmarks: QC-FC correlations, QC-FC distance-dependence, HLM contrasts, tDOF-loss, and TRT benchmarks. We also examined the relative sensitivity of each method in uncovering clinical group differences in two separate case-control samples. One was characterized by relatively high levels of motion and comprised healthy controls and patients with schizophrenia. The other contained low levels of motion and comprised healthy controls and patients with obsessive-compulsive disorder (OCD). The third dataset contained longitudinal data on healthy controls only and was used to examine TRT.

Our comprehensive assessment revealed that when head motion is high, no single pipeline is completely effective in mitigating its contaminating effects on functional connectivity, but that ICA-AROMA (Pruim et al., 2015b; 2015a) consistently shows strong performance across all benchmarks particularly when applied after exclusion of high-motion individuals. We also show that group differences vary dramatically depending on the specific processing pipeline used, even flipping direction in some cases. Our work highlights the need for the comprehensive reporting of motion and its impact on functional connectivity in case-control rs-fMRI studies.

Materials and methods

Participants and data

The rs-fMRI data used in this study were drawn from three sources: (1) the Brain & Mental Health laboratory dataset (*BMH*: 39 HCs and 34 OCD patients); (2) the Consortium for Neuropsychiatric Phenomics dataset (CNP: 121 HCs and 50 schizophrenia patients Poldrack et al., 2016); and (3) the New York University dataset (*NYU:* 29 HCs). The BMH, CNP, and NYU datasets were used to compare the relative performance of the different denoising strategies for removing the effects of in-scanner movement. The BMH and CNP datasets were used to examine which denoising strategy is most sensitive to clinical group differences (relative to HCs) in two disorders: obsessive-compulsive disorder (OCD; provided by BMH) and schizophrenia (SCZ; provided by CNP). By using two independent datasets to achieve

these goals, we sought to establish the generalizability of our findings. The (NYU) dataset (http://fcon_1000.projects.nitrc.org/indi/CoRR/html/), available through the Consortium for Reliability and Reproducibility (CoRR; Zuo et al., 2014), was used to examine within- and between-session test-retest reliability of functional connectivity estimates obtained after the application of each denoising approach.

The BMH dataset was acquired on a Siemens MAGNETOM Skyra 3T scanner. A T1-weighted MP-RAGE structural image was obtained (TE = 2.55 ms, TR = 1.52 s, flip angle = 9°, 208 slices with 1 mm isotropic voxels). Resting state data was obtained using BOLD contrast sensitive gradient echoplanar imaging (EPI) (TE = 30 ms, TR = 2.5 s, flip angle = 90°, 189 volumes, 44 slices). The CNP dataset was acquired on one of two Siemens Trio 3T scanners. A T1-weighted MP-RAGE structural image was obtained (TE = 24 ms, TR = 5 s, flip angle = 90°, 176 slices with 1 mm isotropic voxels). Resting state data was obtained using BOLD-sensitive EPI (TE = 30 ms, TR = 2 s, flip angle = 90°, 152 volumes, 34 slices). The NYU dataset was acquired using a Siemens MAGNETOM Allegra 3T scanner. Details of the T1 scan are TE = 3.25 ms, TR = 2.53 s, flip angle = 7°, 128 slices with 1.3 x 1.0 x 1.3 mm voxels. Details of the EPI scan are TE = 15 ms, TR = 2 s, flip angle = 90°, 180 volumes, 33 slices.

Image processing

EPI and T1-weighted scans were processed using code developed in Matlab, which is freely available through GitHub (https://github.com/lindenmp/rs-fMRI).

Structural image processing

Each participant's T1-weighted high-resolution structural image was processed using the following steps: (1) removing the neck using FSL's *robustfov*; (2) segmentation into WM, CSF, and grey matter (GM) probability maps using SPM8's *New Segment* routine to allow for identification of WM/CSF voxels for use with some pipelines; and (3) nonlinear spatial transform to MNI space using Advanced Normalization Tools (ANTs; Avants et al., 2008) with default settings (using the *antsRegistrationSyN.sh* script).

Core image processing

The functional data underwent a core, common set of processing steps both before and after each denoising method was applied. The core processing pipeline used before denoising included the following steps: (1) removal of the first four volumes of each acquisition; (2) slice-time correction implemented in SPM8; (3) despiking using AFNI's *3dDespike*; (4) two-pass realignment of all volumes to the first volume (first pass) and then to the mean volume (second pass) using SPM8; (5) co-registration of EPI data to the native, cropped, high-resolution structural image via rigid-body registration using ANTs; (6) application of the nonlinear transform derived from the T1-weighted image processing pipeline to the co-registered EPI data using ANTs; (7) linear detrending of the spatially normalized BOLD time series (using *rest_detrend* function from REST toolbox); and (8) intensity normalisation of the EPI data to mode 1000 units. Specific denoising pipelines were then applied at this stage.

After the application of each specific denoising pipeline, additional core processing steps included bandpass filtering between 0.008 and 0.08 Hz using the fast Fourier transform, and spatial smoothing with an 8mm FWHM kernel (although an exception was made for ICA-AROMA, which requires smoothing prior to noise correction, as discussed below). Owing to concerns that performing realignment after despiking (step 3) might lead to inaccurate estimation of the realignment parameters (Power et al., 2015), we repeated our quality-control benchmarks (see below) using realignment parameters obtained before slice-time correction and despiking. We found no difference in our results and report the main findings using realignment parameters obtained after slice-time correction and despiking.

Denoising pipelines

Our primary aim was to comprehensively evaluate the performance of currently popular methods for correcting BOLD time series for the effects of head motion. To this end, we investigated several confound regression strategies. Some denoising pipelines involve a combination of different noise correction strategies. Table 2 outlines the 16 different pipelines analysed here, each containing a particular combination of noise correction methods. Subsequent sections describe each strategy separately.

Table 2. Characteristics of 16 common denoising pipelines analysed here.

Pipeline	Noise corrections methods	No. of regressors
1	6HMP	6
2	6HMP+2Phys	8
3	6HMP+2Phys+GSR	9
4	24HMP	24
5	24HMP+8Phys	32
6	24HMP+8Phys+4GSR	36
7	24HMP+aCompCor	34
8	24HMP+aCompCor50	24+*k*
9	24HMP+aCompCor+4GSR	38
10	24HMP+aCompCor50+4GSR	28+*k*
11	12HMP+aCompCor (Muschelli et al., 2014)	22
12	12HMP+aCompCor50 (Muschelli et al., 2014)	12+*k*
13	ICA-AROMA+2Phys (Pruim et al., 2015b)	2+*k*
14	ICA-AROMA+2P+GSR	3+*k*
15	ICA-AROMA+8Phys	8+*k*
16	ICA-AROMA+8P+4GSR	12+*k*

Notes. HMP, head motion parameters. Phys, average white matter and cerebrospinal fluid signals. GSR, global signal regression. aCompCor, anatomical component correction. *k* denotes an arbitrary number of additional regressors estimated automatically by the denoising method and which can vary from person to person.

Regression of head motion parameters

For each participant, the two-pass realignment of the BOLD data during core image processing yielded six time-series describing in-scanner movement along six dimensions – three translational axes of x, y, and z, and three rotational axes of pitch, roll, and yaw. BOLD time series were regressed against these head motion parameters and various expansion terms. In the 6HMP model, only the original six head motion parameters were employed as covariates. The 12HMP model employed these 6 parameters, as well as expansion terms derived by computing the temporal derivatives of each parameter (calculated as first-order backwards differences in the head motion time series data). The 24HMP model employs these 12 parameters, as well as the squares of both the original and derivative time series (Satterthwaite et al., 2013).

Regression of mean white matter and cerebrospinal fluid signals

A popular method for controlling for additional head motion effects beyond the 6HMP, 12HMP, and 24HMP models, in addition to capturing physiological fluctuations of non-neuronal origin, is to generate a representative time series from tissue compartments that do not include grey matter. This involves extracting an averaged time series from all WM voxels, and separately for all CSF voxels. To this end, we generated binarized tissue masks for WM and CSF after thresholding their respective tissue probability maps obtained from the T1-weighted processing pipeline at 99%. We then generated a binary mask of GM by thresholding the corresponding tissue probability map at 1% and subtracted this mask from the WM and CSF masks. This procedure resulted in a conservative estimate of the WM and CSF volumes, thus ensuring minimal overlap with GM voxels. The resulting averaged WM and CSF signals were then incorporated into denoising procedures either in their original form (2Phys), or along with their temporal derivatives, squares, and squares of derivatives, resulting in an expansion to 8 nuisance signals (8Phys).

Global signal regression

A controversial step in fMRI denoising is global signal regression (GSR), which involves regressing voxel-wise fMRI time series against an averaged signal computed across the entire brain (i.e., across GM, WM and CSF tissue compartments) (Fox et al., 2007; Murphy et al., 2009). We calculated the global signal by averaging across all voxels in the BOLD data using participant-specific masks that covered the entire brain. These masks were generated by taking the union of two whole brain masks created using FSL's *bet* function applied to the spatially normalised EPI and T1-weighted images created during pre-processing. As above, global signal regression was performed by either removing, via linear regression, just the global signal (GSR) or by also removing the temporal derivative, square term, and square of the derivatives to account for higher order effects, totalling 4 nuisance covariates (4GSR).

aCompCor

A popular method for modelling noise in BOLD data is to apply temporal principal component analysis (PCA) to putative nuisance signals (Behzadi et al., 2007; Muschelli et al., 2014). This approach, most commonly embodied by the aCompCor procedure, involves extracting orthogonal components of temporal variance from voxel-wise time series separately for the WM and CSF tissue compartments. Compared to averaging across voxels, PCA offers the

advantage of identifying multiple orthogonal sources of variance in the data, which may better characterise the noise signals present in the WM and CSF tissue compartments.

We defined WM and CSF masks as outlined above. PCA was performed in the time domain on the voxel time series separately for the WM and CSF masks. As per previous work by Muschelli et al. (2014), aCompCor was run using two models that differed in the number of principal components (PCs) extracted. In the aCompCor model, we extracted the leading 5 PCs for each tissue type, yielding 10 confound regressors. In the aCompCor50 models, we extracted the number of PCs that cumulatively explained at least 50% of the variance for each tissue type, yielding a variable number of confound regressors for each participant and each tissue type. The analysis of Muschelli et al. (2014) suggests that aCompCor50 outperforms aCompCor in terms of mitigating motion-related artefacts and increases the specificity of functional connectome estimations. However, aCompCor50 has the potential to cost many more temporal degrees of freedom (see *Loss of temporal degrees of freedom* below).

ICA-AROMA

Recently, an independent component analysis-based tool called ICA-AROMA was introduced by Pruim et al. (2015b; 2015a). This method attempts to automatically identify and remove motion-related artefacts from BOLD data by using FSL's MELODIC (Beckmann et al., 2005) to first decompose the BOLD data into spatially independent components (IC) before applying a predetermined, theoretically motivated classifier to identify ICs as noise or signal. Specifically, ICs were classified as motion-related if any of the following criteria were true: (1) more than 10% of IC voxels were located within CSF; (2) IC time series contained more than 35% high-frequency content, where high-frequency is defined as a fraction of the Nyquist frequency at which higher frequencies explain 50% of the total power between 0.01 Hz and the Nyquist frequency; and (3) ICs exceed a decision boundary from a two-dimensional linear discriminant analysis (LDA) classifier. For the classifier, a two-dimensional feature space was defined using (1) the proportion of IC voxels that overlapped the edges of the brain; and (2) the maximum absolute correlation between the IC time series and parameters in an expanded head motion model, which included 72 motion realignment parameters comprising the same as those in our 24HMP model plus 24 parameters from a single time-point back and forward. Pruim then defined a hyperplane using an LDA classifier trained on manually labelled ICs (labelled as either 'Motion', 'Resting State Network', or 'Other') from 30 rs-fMRI datasets.

Like all the other methods mentioned thus far, ICA-AROMA is applied to each participant's BOLD data separately. Due to the use of a person-specific classifier to set

thresholds for detecting noise components, the number of confound regressors removed from the BOLD data can vary across participants. Unlike the other methods, ICA-AROMA requires data to be spatially smoothed before noise correction. As such, for all pipelines that included ICA-AROMA (see Table 2), spatial smoothing was performed immediately before noise correction rather than after bandpass filtering.

An alternative ICA-based approach is FMRIB's ICA-based X-noiseifier (ICA-FIX) (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). ICA-FIX's implementation involves a much more extensive set of features that requires re-training and manual re-labelling of noise components on each new dataset. Here we focus only on methods that can be applied automatically to rs-fMRI data with minimal user input.

Volume censoring

Apart from the confound regression strategies outlined above, an additional method for addressing head motion confounds in BOLD data involves censoring contaminated data points (i.e., entire volumes). The two techniques investigated here are spike regression (Satterthwaite et al., 2013) and scrubbing (Power et al., 2012). The difference between the two methods lies in how volumes are marked as contaminated and how contaminated volumes are censored (see below). Previous research has found that both spike regression and scrubbing improve data denoising, as quantified by reduced QC-FC distance-dependence and HLM contrasts (Power et al., 2012; Satterthwaite et al., 2013).

For spike regression and scrubbing, volumes were marked as contaminated by thresholding a given participant's FD. There are several different ways of calculating FD (see Yan et al., 2013a for a summary and definition of each), which all correlate with each other at $r > 0.90$, suggesting that they characterise similar properties of the data (Power et al., 2015). However, the scales of the measure differ; thus, the choice of an FD threshold depends on the specific method used to calculate FD. Here, we sought to maintain consistency with past work. We thus used different FD measures for spike regression and scrubbing, following the methodology of Satterthwaite et al. (2013) for the former and Power et al. (2013) for the latter. Specifically, for spike regression, we calculated FD using the root mean squared volume-to-volume displacement of all brain voxels measured from the six head motion parameters, a method sometimes called FD_{Jenk} (Jenkinson et al., 2002; Satterthwaite et al., 2013; Yan et al., 2013a). Based on previous work by Satterthwaite et al. (2013), volumes were marked as contaminated if FD_{Jenk} >0.25mm. Satterthwaite et al. (2013) demonstrated that a single spike regressor placed at the contaminated time point outperformed a box-car function (i.e., marking

a contiguous set of time points as contaminated) that also covered proximal time points. Thus, for each contaminated volume, a separate nuisance regressor was generated that matched the length of the BOLD time-series data, containing a value of 1 at the location of the contaminated volume and 0 elsewhere. As with ICA-AROMA, the number of regressors generated with spike regression varies over participants.

For scrubbing, we calculated FD as the sum of absolute differences in volume-to-volume changes in the six head motion parameters, FD_{Power} (Power et al., 2012; Yan et al., 2013a). Following previous work by Power et al. (2013), volumes were marked as contaminated using a combination of FD_{Power} and BOLD data variance (DVARS). DVARS is an estimate of framewise changes in BOLD signal intensity, and is calculated as the root-mean-squared variance of the temporal derivative of all brain voxel time series (Power et al., 2012). Based on Power et al. (2013), volumes were marked as contaminated if either of the following were true: FD_{Power} >0.2mm or DVARS >0.3% (Power et al., 2013). Thus, immediately before statistical analysis, contaminated volumes were removed from the BOLD data.

Volume censoring involves a costly trade-off, such that every censored volume improves data quality at the expense of data quantity. Removing too many volumes can result in insufficient data to produce reliable estimates of functional connectivity. Consistent with previous literature, we excluded participants with <4 minutes of data after either spike regression or scrubbing (Satterthwaite et al., 2013; Van Dijk et al., 2012).

Network construction

After pre-processing, we generated functional connectivity networks using two independent parcellations, one containing 333 cortical regions (Gordon et al., 2016) and the other containing 264 cortical and subcortical regions (Power et al., 2011). For each region in each parcellation, we estimated the mean time series across all voxels comprising the region, multiplying each voxel time series by its grey matter probability first to weight the time series mean. We then estimated Pearson correlation coefficients between each pair of regional averaged time series. These correlations can be represented as a network, where edges connecting pairs of nodes (brain regions) represent correlation coefficients between resting state fMRI time series. These networks underwent Fisher's *r*-to-*z* transformation to normalise the correlation distribution and facilitate group comparisons.

For the BMH data, eight ROIs from the Power parcellation were discarded from analyses due to poor overlap with participant EPI data (no ROIs were excluded for the Gordon parcellation). For the CNP dataset, five and seven ROIs from the Gordon and Power

parcellations, respectively, were discarded. For the NYU dataset, no ROIs were discarded from either the Gordon or Power parcellations.

We found that our results were qualitatively very similar for both parcellations. We therefore present results obtained with the Gordon parcellation in the main text (see Results) and the Power results in the supplementary material.

Outcome measures

Ideally, noise correction should remove any statistical relationship between functional connectivity and in-scanner movement. A commonly used summary statistic for quantifying the degree of person-specific head motion is the temporal mean of their FD time series (mFD). Here we use FD_{Jenk} to calculate mFD (similar results are obtained using FD_{Power}). This metric was used to evaluate pipeline efficacy by assessing: (1) how the correlation between mFD and functional connectivity (FC), computed across participants for each network edge, changes following denoising (QC-FC); and (2) how this relationship varies as a function of distance between nodes, given prior evidence that movement has a more pronounced effect on the QC-FC correlation for short-range connections (Power et al., 2012; Van Dijk et al., 2012). In addition, we examined four other performance metrics: (1) the change in temporal degrees of freedom caused by each denoising pipeline (tDOF-loss); (2) the sensitivity of each pipeline to differences in functional connectivity between high- and low-motion HCs; (3) the test-retest reliability of each pipeline (TRT); and (4) the sensitivity of each pipeline to clinical differences in functional connectivity. These benchmarks are outlined in more detail in the following.

QC-FC correlations

We used Pearson correlations to quantify the association between subject-specific mFD and functional connectivity estimates for each edge, resulting in an edge-specific QC-FC value estimated across the entire sample. We compared the proportion of edges where this QC-FC correlation was statistically significant ($p<0.05$, FDR corrected) after applying each denoising pipeline as a measure of the efficacy of each approach in removing motion-related variance.

QC-FC distance dependence

In-scanner movement tends to spuriously inflate short-range functional connectivity relative to medium- and long-range connectivity (Power et al., 2012; Van Dijk et al., 2012). This spatial variation arises because each voxel is more strongly contaminated by proximal (vs distal) voxels when the head moves in the scanner, thus spuriously increasing synchrony between

nearby brain regions. Successful noise correction should thus result in no distance-dependence of QC-FC correlations. It has previously been shown that ICA-AROMA is effective at mitigating this relationship while simple linear regression methods are not and GSR exacerbates it (Ciric et al., 2017). Here, we estimated the distance between regions as the Euclidean distance between the stereotaxic coordinates of the volumetric centres of brain region pairs. For each edge, we then quantified the association between this distance measure and the QC-FC correlation for that edge using Spearman's rank correlation coefficient, ρ, due to the non-linearity of some associations. To visualise the relationship between QC-FC and distance, we also plotted the QC-FC correlation of each edge as a function of this distance metric. Since each network contained tens of thousands of edges, the resulting scatterplots displayed a dense point cloud which masked mean-level trends. To provide a more interpretable visualisation of QC-FC distance-dependence, we divided the data into 10 equiprobable bins (using equally spaced quantiles to define bins) based on nodal distance, and plotted the mean and standard deviation of QC-FC correlations in each bin.

Loss of temporal degrees of freedom

The pipelines examined here vary in terms of the number of regressors used to model noise in the fMRI time series. Using more nuisance regressors can capture additional sources of noise-related variance in the data and thus improve denoising, but this comes at the expense of a loss of temporal degrees of freedom. The number of time points in a BOLD dataset represents the degrees of freedom available for statistical inference, and fewer degrees of freedom can spuriously increase functional connectivity (Yan et al., 2013a). Furthermore, noise correction strategies that use variable degrees of freedom across participants (e.g., volume censoring and ICA-AROMA) can lead to artefactual group differences in functional connectivity (Pruim et al., 2015b; 2015a; Yan et al., 2013b). Thus, the relative performance of different denoising strategies must be balanced against lost temporal degrees of freedom (tDOF-loss). For each pipeline, we calculated tDOF-loss as the number of regressors and, in the case of volume censoring, the number of contaminated volumes.

Group differences in high-motion and low-motion healthy participants

Apart from mitigating QC-FC correlations, successful noise correction should also yield minimal group differences in functional connectivity between HC participants that differ on their amount of in-scanner movement. Thus, for the BMH and CNP datasets, we split HC

participants into three equally sized groups based on mFD and mapped group-differences in functional connectivity using mass-univariate statistics. Specifically, for each edge we calculated *t*-contrasts assessing both increases and decreases in functional connectivity in high-motion HCs compared to low-motion HCs (HLM contrasts; medium-motion group was not included). Age (demeaned) and sex were entered as covariates, and in the case of the CNP dataset, scanner site was also included as a nuisance covariate. Furthermore, denoising pipelines that yielded variable tDOF-loss across participants, tDOF-loss (demeaned) was included as a nuisance covariate. We then thresholded the *t*-contrasts at $p<0.05$ FDR-corrected and $p<0.05$ uncorrected and report the proportion of significant edges for each pipeline.

Test-retest reliability

A good denoising strategy will yield consistent and reliable estimates of functional connectivity across repeated measurements of the same subject under the same conditions. Previous work has shown that the TRT reliability of functional connectivity decreases with increasing variance explained by denoising models (Birn et al., 2014; But see Van Dijk et al., 2012; Yan et al., 2013a). This may be due to the motion artefact present in BOLD having its own moderate TRT reliability, which, if not properly removed, may increase TRT reliability of functional connectivity (Yan et al., 2013a). To characterize the effects of noise correction pipelines on the reliability of functional connectivity estimates, we examined TRT reliability using the intra-class correlation (ICC) coefficient (Shrout and Fleiss, 1979) calculated on the NYU dataset, defined as

$$ICC = \frac{MS_b - MS_w}{MS_b + (n-1)MS_w},$$

where MS_b is the between-subject mean square for each edge, MS_w is the within-subject mean square for each edge, and n is the number of observations per participant (here n is always to set to 2 since there is only ever one repeated scan). To examine both within- and between-session ICC, we retained participants from the NYU dataset that had three rs-fMRI scans obtained across two scan sessions. Scans 1 and 2 were collected during session 1 and were separated within session by 28 minutes, on average (SD = 9 days). Scan 3 was collected in session 2, which was separated from session 1 by an average of 90 days (SD = 72 days). Intra-session ICC was computed using scan 1 and scan 2 and inter-session ICC using scan 1 and scan 3.

Sensitivity to clinical differences in functional connectivity

Resting-state fMRI is widely used to characterize case-control differences in functional connectivity. Motion can be a major confound in these analyses, particularly as some patient populations may be more prone to move in the scanner, for example due to heightened anxiety in the MRI environment, medication side-effects, or symptoms of the disease itself (e.g., tremor in Parkinson disease). An optimal denoising pipeline will remove spurious differences in functional connectivity that are driven by motion, and therefore isolate any "true" differences between groups. We assessed how different pipelines impact sensitivity to detect clinical differences by conducting case-control comparisons in: (1) the BMH dataset, comparing HCs and OCD patients; and (2) the CNP dataset, comparing HCs and SCZ patients.

To keep our analyses comparable with the wider literature examining pathophysiological mechanisms in mental health, we evaluated sensitivity to clinical group differences using the network-based statistic (NBS; Zalesky et al., 2010). The NBS improves statistical power over typical mass-univariate correction strategies such as the FDR by testing the null hypothesis at the level of connected components of edges (rather than individual edges). These components are formed by applying an initial threshold to the data of $p<0.05$ uncorrected. The observed component sizes are compared to an empirical null distribution of maximal component sizes obtained by permuting group membership 5,000 times. Evaluating observed component sizes with respect to a null distribution of maximal sizes ensures that the resulting inferences on network components are corrected for multiple comparisons ($p<0.05$, corrected component-wide; see Fornito et al., 2016). NBS *t*-contrasts were performed in both directions, assessing both increases and decreases in functional connectivity in patients compared to HCs. As with HLM contrasts, age (demeaned), sex, and if applicable, scanner site and tDOF-loss were entered as nuisance covariates. For each significant sub-network of edges identified by the NBS, we quantified its size as the proportion of significant edges in the subnetwork relative to the total number of connections in the parcellation.

Results

Sample characteristics

Following Satterthwaite et al. (2013) we excluded one participant from the CNP data with high levels of gross motion (>0.55 mFD). No participants were removed from the BMH or NYU datasets using this criterion.

At the sample-level, average mFD was 0.07±0.03 (mean±SD) and 0.10±0.06 for the BMH and CNP datasets. For the NYU dataset, average mFD was 0.06±0.05 for scan 1 and 0.06±0.03 for scans 2 and 3. The mFD was significantly higher in the CNP dataset relative to both the BMH and NYU datasets (BMH: $t = 5.90$, $p<0.001$; NYU scan 1: $t = 3.26$, $p = 0.0013$), indicating that participants in the CNP dataset had greater in-scanner movement (mFD did not differ between BMH and NYU).

For the CNP data, average mFD was significantly lower in HC (0.09±0.05) compared to SCZ (0.13±0.06; $t(84) = 5.12$, $p<0.001$). for the BMH data, average mFD was virtually identical for HC and OCD groups from the BMH dataset (both 0.07±0.03). Thus, motion was higher, on average, in the CNP compared to BMH data and, within each dataset, there were differences between patients and controls in the CNP but not BMH data.

QC-FC correlations

Our first aim was to examine the efficacy of different noise correction pipelines in removing the relationship between participant in-scanner movement and estimates of functional connectivity, as assessed with QC-FC correlations. Separately for the BMH, CNP and NYU datasets, we collapsed patients and HCs into a single group to compute this correlation. Figure 1 shows the percentage of 55,278 edges from the Gordon parcellation with significant QC-FC correlations ($p<0.05$, FDR-corrected. See Fig. S1 for results using $p<0.05$, uncorrected) as well as the distributions of QC-FC correlations, for the BMH, CNP, and NYU datasets, following application of each of the 16 denoising pipelines listed in Table 2.

For the low-motion BMH data, the proportion of edges showing significant QC-FC correlations ($p<0.05$, FDR-corrected) was <1% for all pipelines except 12HMP+aCompCor and 24HMP+aCompCor, where the proportion of edges showing significant QC-FC correlations was 11-12%. In the NYU data, which was also characterized by low motion, no significant QC-FC correlations were identified at the FDR-corrected threshold, possibly due to the smaller sample size of the NYU data. At sub-FDR threshold, the QC-FC distributions

showed a similar profile across pipelines to the BMH data, with elevations for the 12HMP+aCompCor and 24HMP+aCompCor pipelines, in addition to 6HMP and 24HMP (Fig. 1C). These findings suggest that, in low-motion data, aCompCor without GSR may compound motion-related artefact.

No pipelines successfully reduced QC-FC to 0% in the high-motion CNP dataset. Simple strategies relying on motion parameters alone (i.e., 6HMP and 24HMP) fared the worst, with 58% and 68% of edges showing significant QC-FC correlations, respectively ($p<0.05$, FDR-corrected). This result suggests that most measures of functional connectivity in analyses using HMP correction alone are contaminated by motion, as similarly reported by Ciric et al. (2017). Adding even simple tissue-averaged physiological regressors (e.g., 2Phys or 8Phys, with or without GSR) substantially improved the performance of the HMP models, reducing the number of significant QC-FC correlations by ~1/3.

Replacing tissue-averaged estimates of physiological noise with aCompCor regressors did not improve the performance of the HMP models, with the 12HMP+aCompCor and 24HMP+aCompCor pipelines showing significant QC-FC correlations in around 50% of edges. GSR dramatically improved the performance of aCompCor, with only 7.3% of edges showing residual effects of motion in the 24HMP+aCompCor+4GSR pipeline. Using more principal components also improved performance (12HMP+aCompCor50 and 24HMP+aCompCor50). The best performance across all pipelines was obtained with the 24HMP+aCompCor50+4GSR pipeline, with only 2.7% of edges showing residual QC-FC correlations.

Pipelines incorporating ICA-AROMA generally performed well, with <10% of edges showing significant residual effects of motion. The best performing pipeline was ICA-AROMA+2Phys, with 7.1% of edges showing significant QC-FC correlations. Adding GSR to the ICA-AROMA pipeline led to a slight decrement in performance, with 9.1% of edges showing a significant QC-FC correlation.

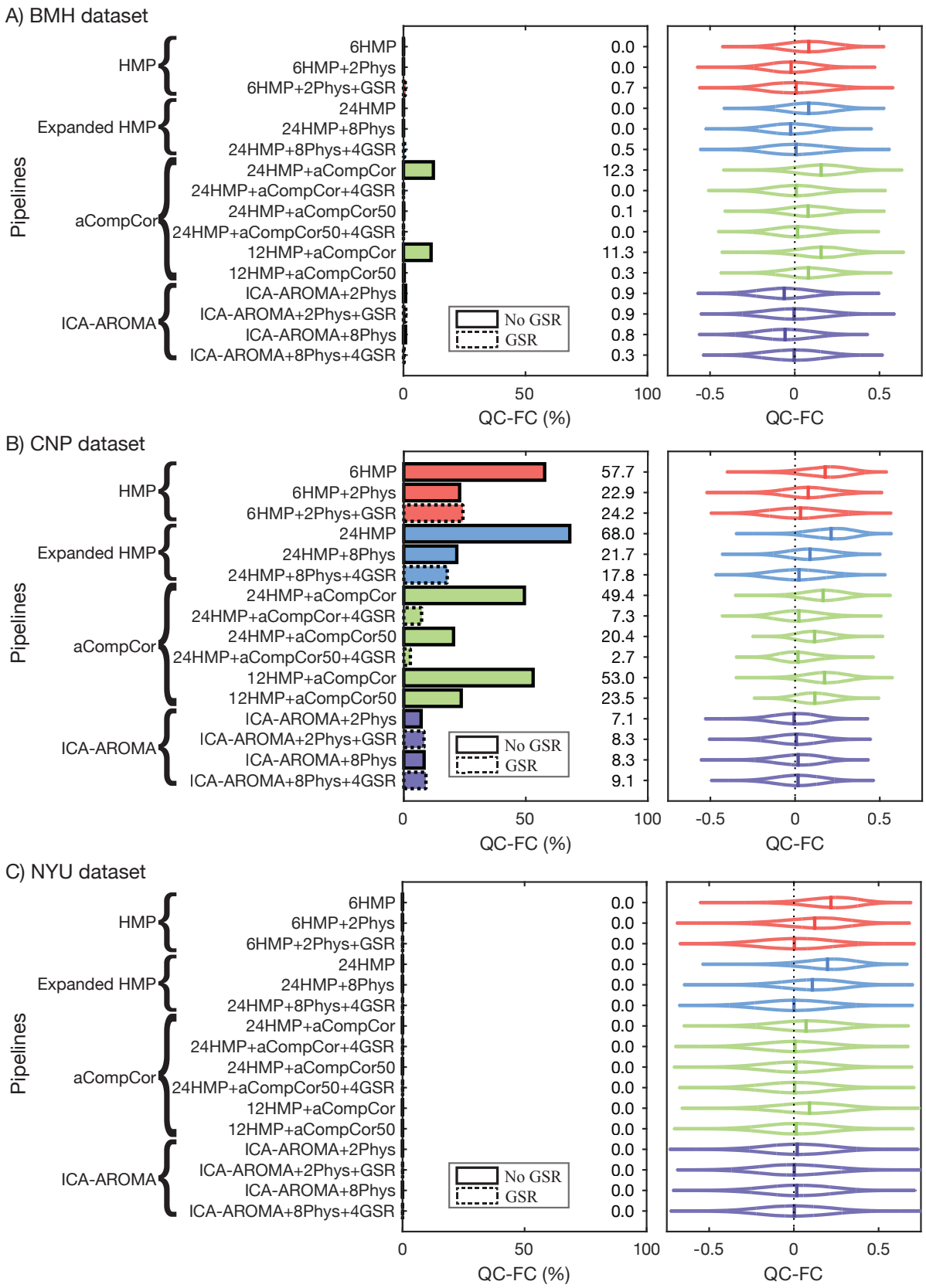

Figure 1. The residual effect of in-scanner motion on functional connectivity after noise correction with one of sixteen different rfMRI pre-processing pipelines. Functional connectivity at each edge was correlated with a summary metric of in-scanner movement across the entire sample (QC-FC correlations) for three separate datasets. The proportion of

functional connections that correlated significantly ($p<0.05$, FDR-corrected) with subject motion, as well as the full distributions of QC-FC values, are shown for each pipeline for the BMH (*A*), CNP (*B*), and NYU (*C*) datasets separately. Dotted lines around the horizontal bars denote pipelines that incorporated global signal regression (GSR).

QC-FC distance dependence

Head motion can spuriously inflate short-range coupling and attenuate long-range connectivity. An ideal denoising procedure should leave no residual association between QC-FC and inter-regional distance. Figure 2 shows the Spearman rank correlation coefficient between QC-FC correlations and edge-wise Euclidean distance for each pipeline. Figure 3 visualises this relationship across a series of equiprobable distance bins for the high-motion CNP dataset. Visualisations for the low-motion BMH and NYU datasets appear in Figures S2 and S3.

For the CNP dataset, ICA-AROMA pipelines generally showed the weakest distance dependence, particularly when combined with GSR/4GSR. The aCompCor/aCompCor50 and HMP+Phys pipelines were slightly worse, and similar to each other. A similar trend was evident in the BMH data. Distance-dependence was low for all pipelines in the NYU dataset.

For methods that were not based on ICA-AROMA, GSR/4GSR generally increased distance-dependence in all datasets. This effect occurred despite GSR having a generally positive effect on QC-FC correlations (Fig. 1). For instance, the 24HMP+aCompCor50+4GSR pipeline, which was most effective at reducing QC-FC correlations for the CNP dataset (Fig. 1B, QC-FC = 2.7%), showed one of the highest levels of distance-dependence (Fig. 2B). This result fits with previous reports demonstrating that GSR improves QC-FC correlations at the cost of increased distance-dependence (Ciric et al., 2017). Taken together, these results suggest that ICA-AROMA performs well at reducing QC-FC correlations, and is particularly effective at mitigating distance-dependence.

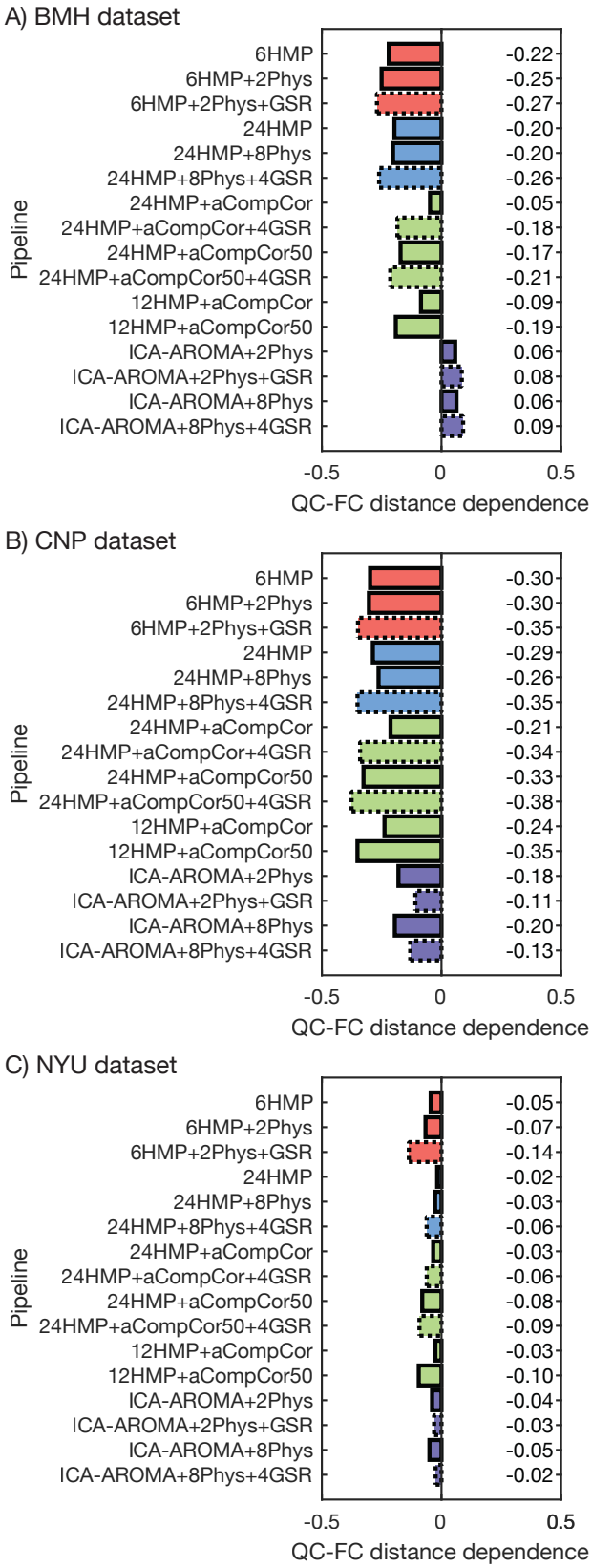

Figure 2. The distance-dependence of QC-FC correlations following application of each pipeline to each dataset. Distance-dependence was quantified here as a Spearman rank correlation between the QC-FC correlation of a given edge and the Euclidean distance separating the two coupled regions, estimated separately for the BMH (*A*), CNP (*B*), and NYU (*C*) datasets.

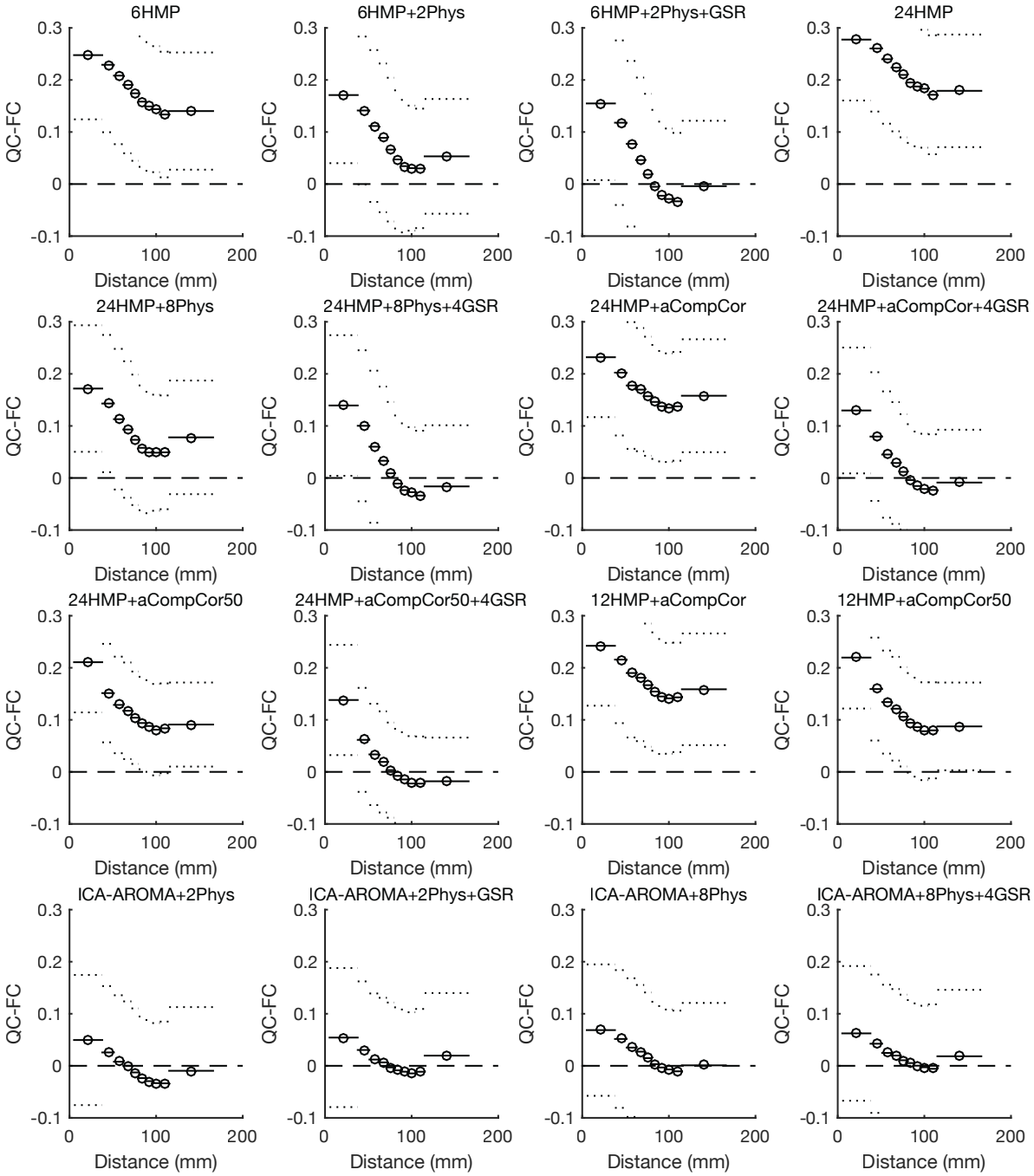

Figure 3. QC-FC distance dependence for each pipeline applied to the CNP dataset. QC-FC correlations were split into 10 equiprobable bins (using equally spaced quantiles to define bins) based on the distance between nodes. For each bin, the mean distance between edges (circle) as well as the mean (solid line) and standard deviation (dotted line) of QC-FC correlation are shown.

Volume censoring

We next investigated the effect of volume censoring on a subset of our top performing pipelines in the BMH and CNP datasets. Specifically, we examined the effects of spike regression and scrubbing on the 24HMP+8Phys+4GSR, 24HMP+aCompCor+4GSR and ICA-AROMA+2Phys pipelines. For scrubbing, 7 and 0 participants were excluded from the CNP and BMH datasets, respectively, due to having <4 minutes of BOLD data remaining after censoring. For spike regression, 16 participants were excluded from the CNP dataset, whereas all participants from the BMH sample were retained.

The key results are summarized for the CNP dataset in Figure 4, which shows the QC-FC correlations and QC-FC distance-dependence before excluding participants according to 4-minute rule (i.e., the exact same participants as those analysed in Fig. 1 and Fig. 2), after exclusion but before censoring, and after exclusion and censoring, for each of the three pipelines (Fig. 4A for scrubbing and Fig 4B for spike regression). The Figure shows that exclusion of participants with high numbers of contaminated volumes can dramatically improve QC-FC correlations and QC-FC distance-dependence, and that there is minimal further improvement provided by censoring the remaining data. In other words, it seems that the primary benefit of volume censoring in these data comes from the identification of participants with high levels of motion that should be removed from the analysis, and not from actual censoring of the data itself. The only exception to this rule was in the combination of ICA-AROMA+2Phys with scrubbing, in which censoring resulted in a slight increase of QC-FC correlations. This effect seems unique to the CNP dataset because we have not observed it in other datasets. In general, the criteria for excluding participants applied in the spike regression pipeline were more effective in reducing QC-FC correlations than the exclusion criteria used in the scrubbing pipeline.

The effect of volume censoring on the low-motion BMH dataset was minimal since all QC-FC correlations were already <1%. The largest reduction in QC-FC correlations was 0.3% for the ICA-AROMA+2Phys pipeline using spike regression.

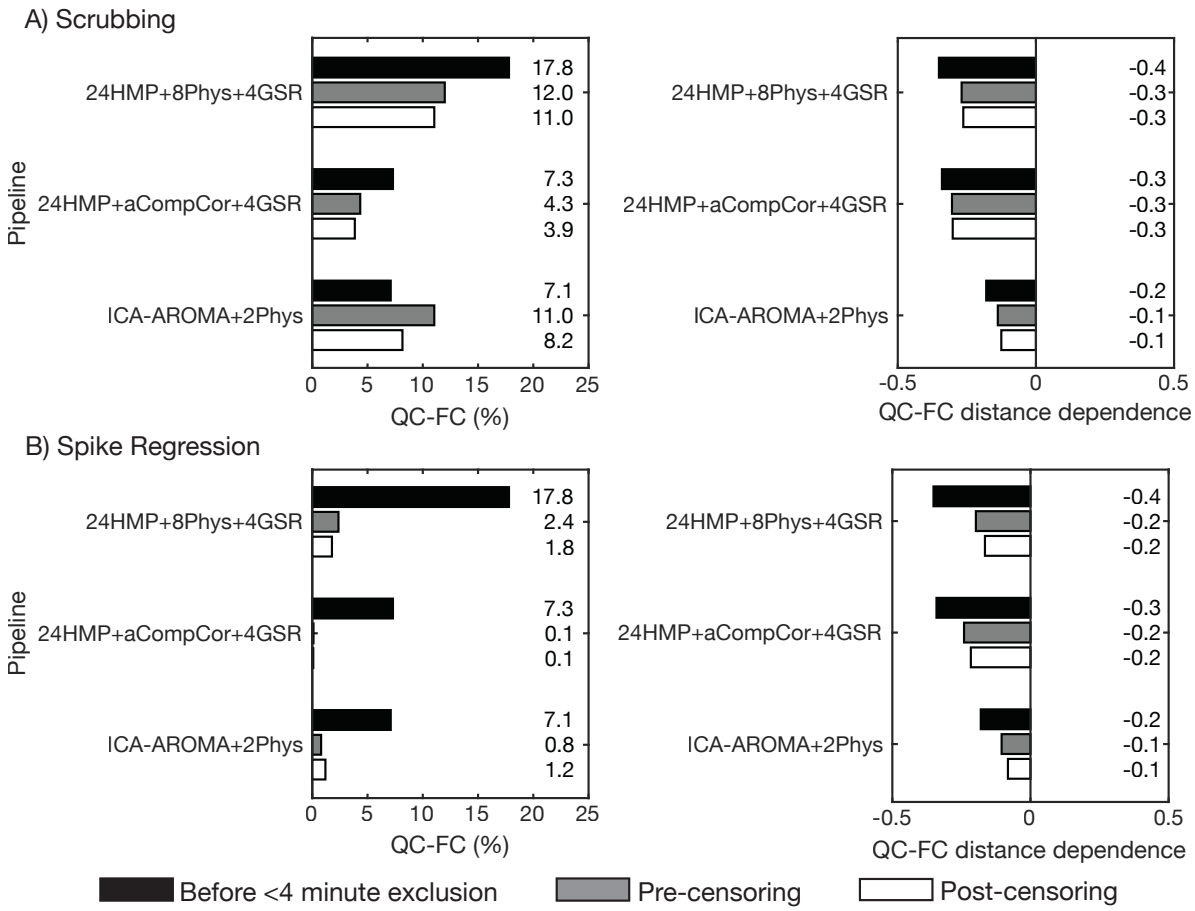

Figure 4. Effect of volume censoring on mitigating motion-related artefact in the CNP dataset. We plot the proportion of edges exhibiting significant QC-FC correlations (left) and QC-FC distance-dependence as Spearman correlation coefficients (right) for scrubbing (A) and spike regression (B). Apart from the combination of ICA-AROMA and scrubbing, the proportion of edges showing significant QC-FC decreases after excluding participants with <4 minutes of non-contaminated data (pre-censoring). Censoring the remaining data provides little further benefit to QC-FC correlations (post-censoring), irrespective of whether censoring is performed via (A) scrubbing, or (B) spike regression. Spearman correlation coefficients are less impacted by exclusion (pre-censoring) and by censoring (post-censoring) compared to QC-FC correlations. The exclusion criteria used for spike regression are more effective in reducing QC-FC correlations than those used for scrubbing.

Loss of temporal degrees of freedom

Using an increasing number of nuisance regressors reduces the degrees of freedom in the BOLD data, which can lead to spurious increases in functional connectivity (Yan et al., 2013a). Thus, we next characterised the loss of temporal degrees of freedom (tDOF-loss) for each pipeline and dataset, as shown in Figure 5. For all datasets, the aCompCor50 pipelines resulted in the highest tDOF-loss, with the loss often exceeding double the tDOF-loss of the other pipelines. As expected, the relatively simple 6HMP class of models yielded the lowest tDOF-loss, with ICA-AROMA and 24HMP pipelines resulting in similar levels of tDOF-loss.

Interestingly, the ICA-AROMA pipelines showed lower tDOF-loss in the high-motion CNP dataset compared to the low-motion BMH dataset, suggesting that ICA-AROMA may be more effective at identifying noise components in high-motion compared to low-motion datasets.

For many of our pipelines, the number of nuisance regressors, and hence the number of degrees of freedom, is explicitly set in the model and invariant across participants. However, for ICA-AROMA and aCompCor50, the temporal degrees of freedom are free to vary across participants. When comparing two or more groups, this variability in tDOF-loss may confound comparisons of functional connectivity (Yan et al., 2013b). To investigate this possibility in our data, we performed two-sample *t*-tests comparing tDOF-loss between HCs and patients in the BMH and CNP datasets for the 24HMP+aCompCor50 and ICA-AROMA+2Phys pipelines. For the BMH dataset, no significant difference was observed in tDOF-loss between HCs and OCD participants for either pipelines. For the CNP dataset, tDOF-loss was significantly higher in the SCZ participants (mean tDOF-loss = 22±6) compared to HCs (mean tDOF-loss = 18±5) for the ICA-AROMA+2Phys pipeline ($t(77) = 4.29$, $p < 0.001$, uncorrected). No difference was observed for the 24HMP+aCompCor50 pipeline.

These results suggest that ICA-AROMA, when combined with exclusion of participants with high levels of motion, is effective at reducing QC-FC and QC-FC distance-dependence for only a moderate tDOF-loss. However, if there are systematic group differences in the amount of head motion, tDOF-loss may also vary across groups when using this approach. Any subsequent analysis must therefore aim to understand the impact of this variation on study findings.

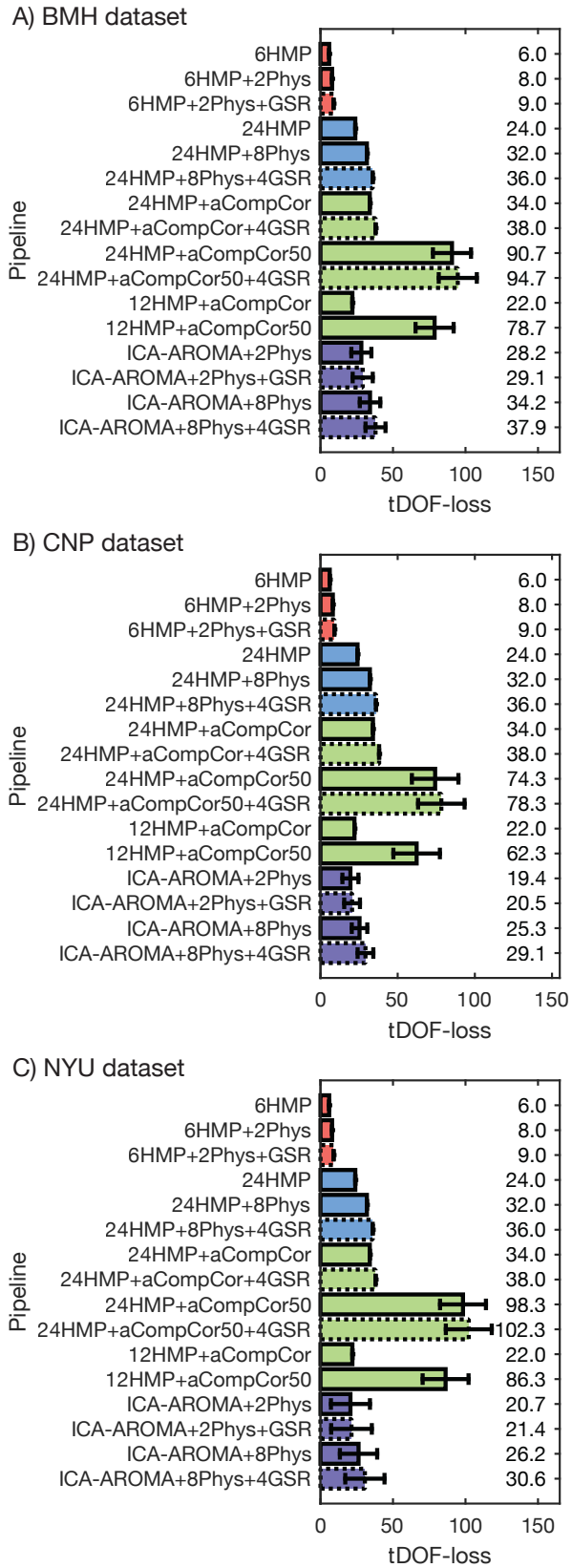

Figure 5. Different fMRI noise correction methods reduce tDOF by different amounts in the (A) BMH, (B) CNP, and (C) NYU data. The ICA-AROMA and aCompCor50 methods define a variable number of noise regressors in different participants. For these pipelines, the sample mean and standard deviation (error bars) are presented. Pipelines using aCompCor50 yield the

highest tDOF-loss. ICA-AROMA pipelines offer relatively modest tDOF-loss despite being highly effective at mitigating motion artefacts.

Group differences in high-motion and low-motion healthy participants (HLM contrasts)

Assuming similar demographic characteristics, two groups of healthy participants should exhibit similar functional connectivity patterns. As such, any differences observed between groups of HCs split by motion can be attributed to ineffective denoising. We thus examined differences in functional connectivity between low-motion and high-motion subsets of HC participants from the BMH and CNP datasets. The NYU dataset was not examined due to an insufficient sample size. Due to the large impact of excluding high-motion individuals (Fig. 4), we performed HLM contrasts after excluding participants who had <4 minutes of data, as identified by spike regression criteria.

Figure 6 shows the sizes of the significant whole-brain effects (represented as a percentage of connections comprising the significant component), identified in a contrast of high-motion HCs and low-motion HCs for each pipeline and dataset. In both datasets, high-motion HCs always showed increased functional connectivity relative to low-motion HCs. However, the pipelines that showed motion-related group differences varied between the BMH and CNP datasets.

In the low-motion BMH dataset, no effects were observed for the FDR-corrected HLM contrasts. In the high-motion CNP dataset, the relative performance of the different pipelines was similar to the analysis of QC-FC correlations (Fig. 1B); that is, pipelines that showed high levels of QC-FC correlations were also more sensitive to connectivity differences between high- and low-motion subgroups. All the HMP+Phys pipelines, except 6HMP+2Phys+GSR and 24HMP+8Phys+4GSR, yielded significant FDR-corrected effects.

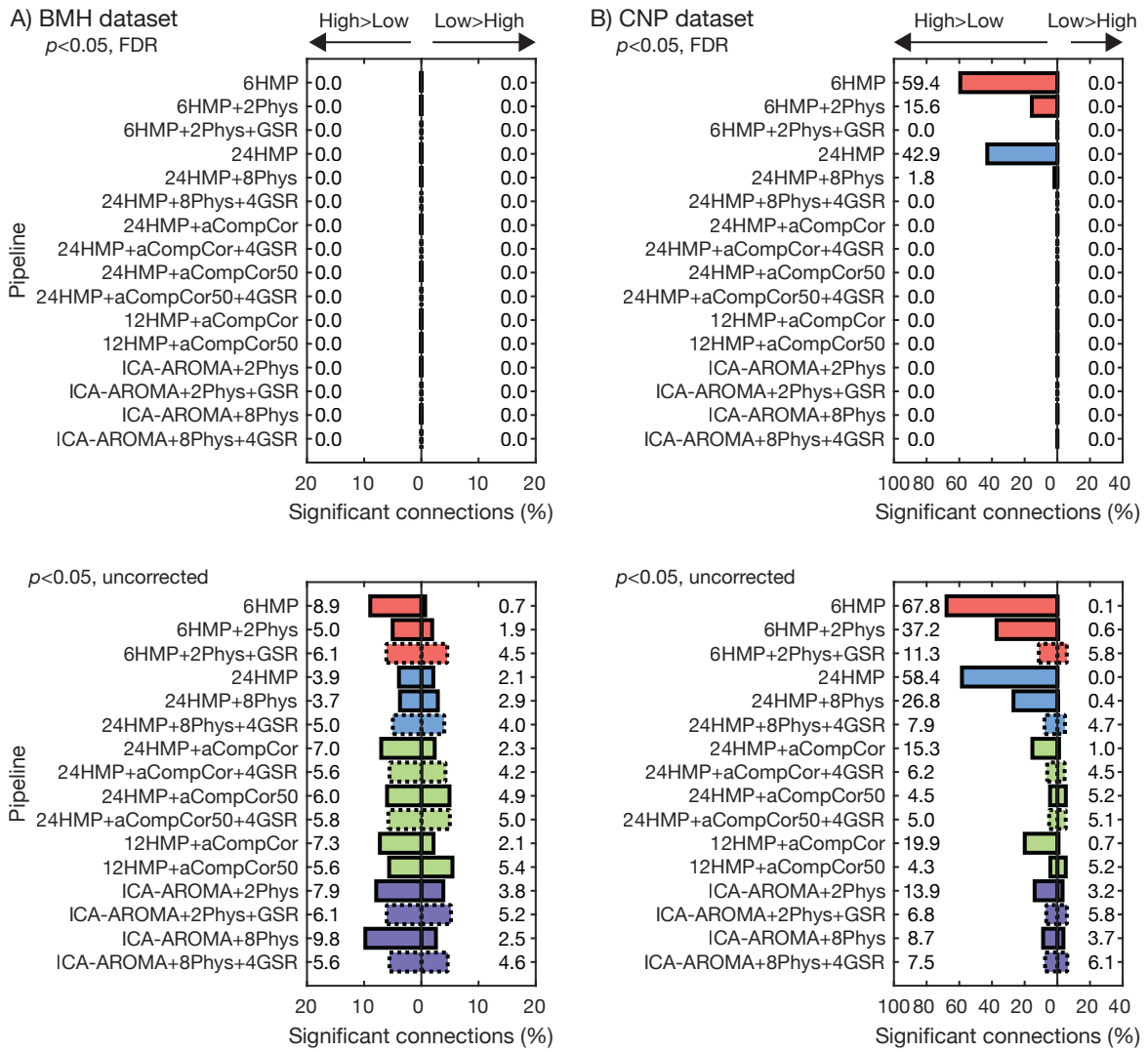

Figure 6. Differences in functional connectivity between groups of healthy controls (HC) that differ on their amount of in-scanner movement. Differences in functional connectivity are plotted as the percentage of edges in which functional connectivity was significantly higher for high-motion compared to low-motion HCs at (top) $p<0.05$, FDR-corrected, and (bottom) $p<0.05$, uncorrected, for the (*A*) BMH and (*B*) CNP datasets.

Test-retest reliability

Next, we examined test-retest reliability across our pipelines using the intraclass correlation coefficient (ICC) calculated using the NYU dataset. High ICC is obtained when estimates of functional connectivity at each edge are consistent across repeated scans. Means and standard deviations of ICC are plotted for each pipeline separately for within- and between-session TRT in Figure 7. In these plots, the pipelines have been ordered by tDOF-loss to reveal a general trend in which ICC decreases as tDOF-loss increases, as shown by previous analysis of different HMP models (Yan et al., 2013a). In other words, the results in Figure 7 show TRT of

functional connectivity decreases as the number of nuisance regressors increases. For instance, when using aCompCor50 pipelines for noise correction, which yield upwards of ~90 nuisance regressors (see Fig. 5), the mean ICC is nearly zero. By comparison, the simplest noise regression model, 6HMP, yielded the highest short-term (within-session) and long-term (between-session) reproducibility. ICA-AROMA+2Phys, which performed well on the other benchmarks, showed relatively high-to-moderate TRT.

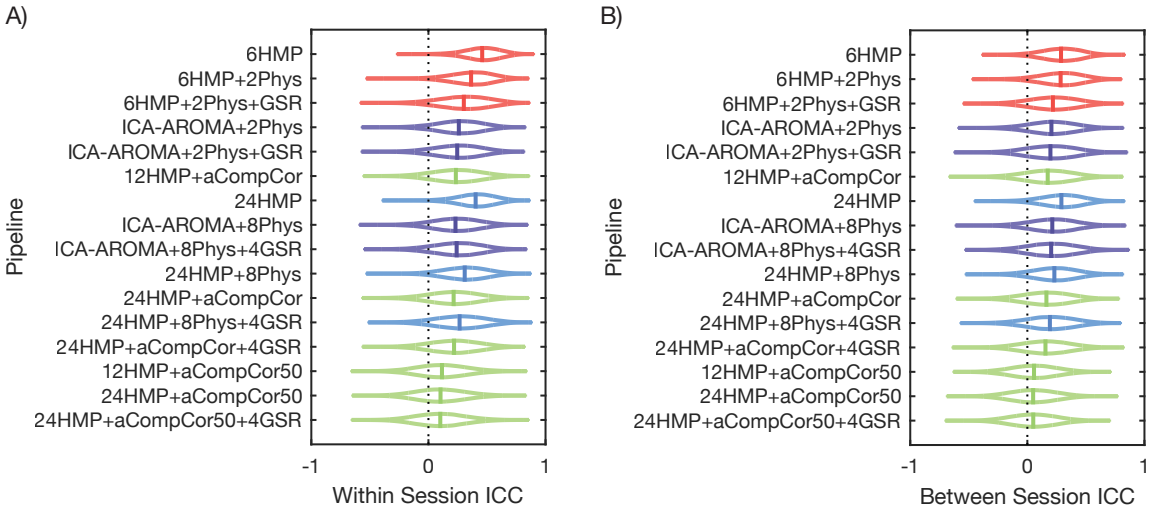

Figure 7. Test-retest reliability varies over fMRI noise correction pipelines as a function of tDOF-loss. (*A*) Intra-class correlation (ICC) coefficients for within session test-retest from the NYU dataset. (*B*) ICC coefficients for between session test-retest from the NYU dataset. Pipelines are ranked on the horizontal axis by descending loss in temporal degrees of freedom (tDOF-loss) calculated using fMRI data from baseline (i.e., session 1, scan 1).

Group differences in functional connectivity

We next examined the sensitivity of each pipeline to clinical group differences in functional connectivity in the BMH and CNP datasets using *t*-contrasts corrected for multiple comparisons via the NBS and FDR-correction. No significant group differences were found using either thresholding approach in the comparison of OCD patients and HCs.

In the comparison of HC and SCZ, the NBS was generally more sensitive to differences than FDR-correction. Only ICA-AROMA+2Phys yielded any significant FDR-corrected results, with 0.5% of edges showing reduced functional connectivity in schizophrenia relative to HC (HC>SCZ). Several different pipelines showed group differences between SCZ patients and HCs using the NBS. For each pipeline and contrast, only one significant NBS component was found. Each significant component is represented as a percentage of all possible edges in Figure 8. As above, participants with <4-minutes of data as per spike regression criteria were excluded from analysis.

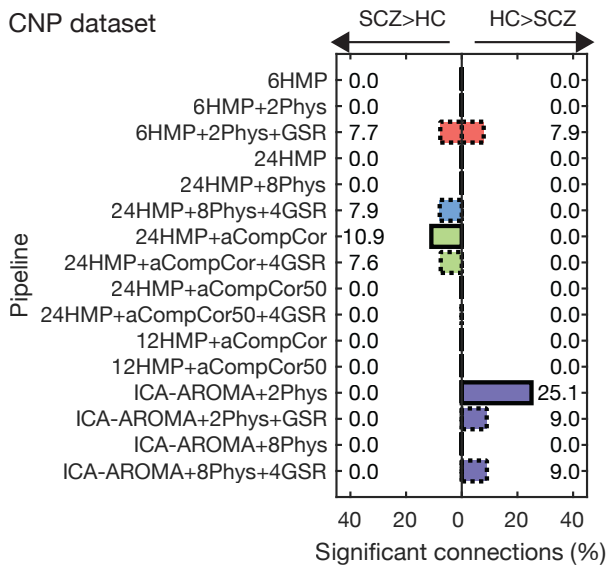

Figure 8. Case-control differences depend on denoising strategy for the CNP dataset. The proportion of edges contained in the NBS subnetwork showing significant differences between healthy controls and individuals with schizophrenia, $p<0.05$, component-wide corrected.

Across all pipelines, the largest difference, comprising ~25% of all possible edges, was observed for the HC>SCZ contrast using the ICA-AROMA+2Phys pipeline. This finding of a large deficit of functional connectivity is consistent with prior evidence for global reductions of functional connectivity in people with schizophrenia (Fornito et al., 2011a; 2012), and evidence of generally reduced structural connectivity in the disorder (Pettersson-Yeo et al., 2011; Zalesky et al., 2011). This result is reassuring, given that the ICA-AROMA+2Phys pipeline was amongst the best performing on all the other benchmarks, suggesting that the efficacy of this denoising procedure in mitigating the effects of motion was coupled with an enhanced sensitivity for detecting group differences of putative pathophysiological significance. Adding GSR to this pipeline reduced the size of the subnetwork to 9% of edges. Using 8Phys with ICA-AROMA resulted in no significant group differences whereas adding 4GSR to ICA-AROMA+8Phys identified an NBS component also comprising 9% of edges.

For the reverse SCZ>HC contrast, the 24HMP+8Phys+4GSR, 24HMP+aCompCor, 24HMP+aCompCor+4GSR showed significant components each comprising ~7-11% of edges. Only one pipeline, 6HMP+2Phys+GSR, showed significant components in both directions, each comprising ~7% of edges. Notably, three of the four pipelines showing differences in the SCZ>HC direction included GSR. This association with GSR was even more striking when using the Power parcellation, where differences in the SCZ>HC direction were also identified when adding GSR to ICA-AROMA pipelines (Fig. S8). These results are

alarming, and indicate that the direction of group differences can flip depending on how the data are pre-processed, as has been suggested previously (Gotts et al., 2013; Saad et al., 2012).

We next sought to understand whether group differences identified by distinct pipelines in the same direction (i.e., HC>SCZ or SCZ>HC) implicated similar sets of edges. In other words, we sought to examine whether pipelines showing differences in the same direction identified the same sets of edges as differing between groups. To address this question, we visualised the number of significant NBS connections that overlapped for each pair of pipelines that identified a difference in the same direction (Fig. 9). Overlap was estimated using the Jaccard Index, which is the number of the intersecting edges in a significant component normalized by the union of the component edges across the two pipelines. A value of 0 indicates no overlap between edges and a maximal value of 1 indicates perfect overlap. As shown in Figure 9, the overlap between pipelines generally quite low, and never exceeds 50%. Higher overlap was generally observed for pipelines incorporating GSR.

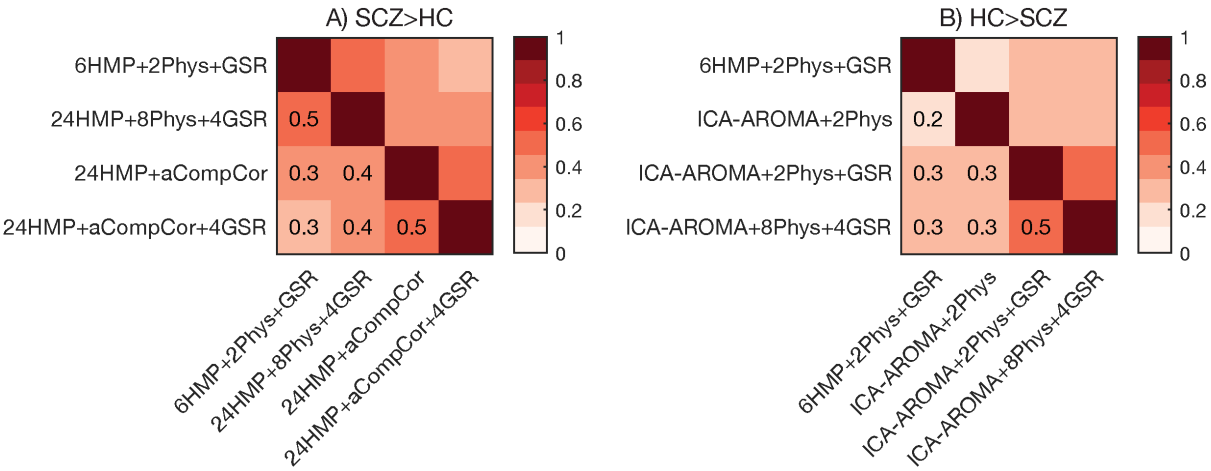

Figure 9. Proportion of connections that overlap between pairs of select pipelines showing significant NBS components for the SCZ>HC (*A*) and HC>SCZ (*B*) contrasts. Proportions are calculated by dividing the intersection of the significant edges by the union of the significant edges. Despite revealing significant NBS components in the same direction, not all pipelines yield NBS components with similar profiles of significant connections.

To examine in detail the implications of this variability for interpreting group differences in functional connectivity, we visualise in Figure 10 the top 200 connections, ranked by *t*-values, for the ICA-AROMA+2Phys, 6HMP+2Phys+GSR, and 24HMP+8Phys+4GSR pipelines, and examined the proportion of connections that fell within and between the canonical brain subnetworks defined in the Gordon parcellation (Fig. 10). We

focused on these three pipelines because ICA-AROMA+2Phys has consistently performed well across our different benchmarks and 6HMP+2Phys+GSR and 24HMP+8Phys+4GSR are commonly used denoising procedures which, in this analysis, revealed differences either in both directions (6HMP+2Phys+GSR) or in the SCZ>HC direction (24HMP+8Phys+4GSR).

Qualitatively, visualizing the anatomical distribution of the group differences (Fig 10, top) suggests that the strongest differences are for edges running in the anterior-posterior direction with the 24HMP+8Phys+4GSR pipeline and for inter-hemispheric edges with the ICA-AROMA+2Phys and 6HMP+2Phys+GSR pipelines. When the differences are broken down by sub-network, we arrive at very different conclusions depending on which pipeline is used. For example, 24HMP+8Phys+4GSR heavily implicates the visual system as showing *increased* functional connectivity in schizophrenia (38% of edges in total), whereas ICA-AROMA+2Phys implicates the default mode network showing *decreased* functional connectivity in schizophrenia (27% of edges in total). By comparison, 6HMP+2Phys+GSR identifies increased functional connectivity in patients mainly in visual areas and decreased functional connectivity mainly areas that have not been clearly assigned to a network in the Gordon parcellation.

Together, these results indicate that the choice of noise correction strategy can have a profound effect on case-control comparisons. In the high-motion CNP dataset, the functional dysconnectivity in schizophrenia, as measured by rs-fMRI, flipped direction depending on which pipeline was used. Even when group differences were identified in the same direction, the specific sets of edges identified as differing between groups showed considerable variation, leading to very different conclusions about which neural systems are affected by disease.

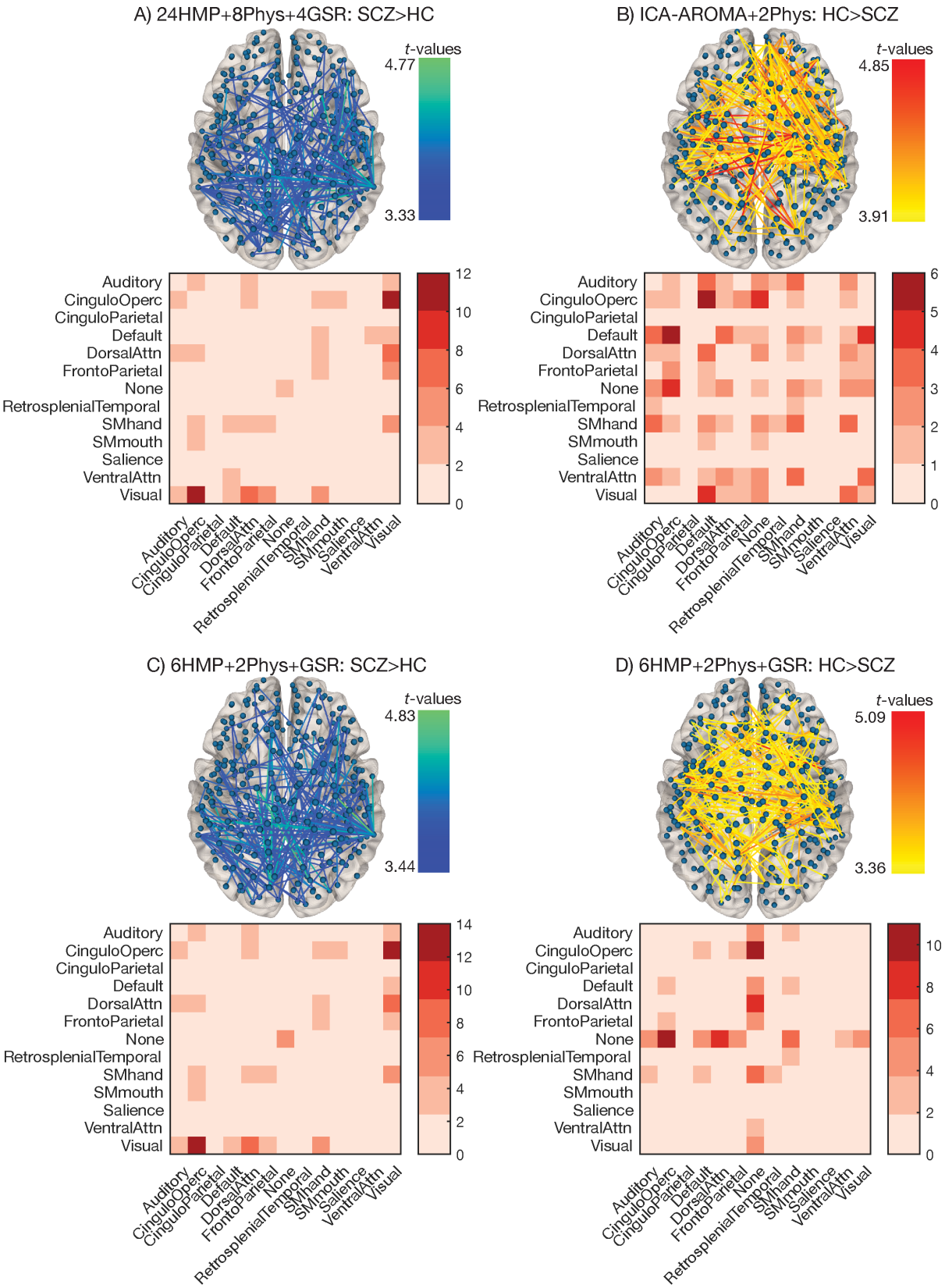

Figure 10. Denoising pipelines impact the direction and anatomical distribution of group differences in functional connectivity. Here we summarize functional connectivity differences

between patients with schizophrenia and healthy control participants in the CNP dataset as characterised using the (*A*) 24HMP+8Phys+4GSR, (*B*) ICA-AROMA+2Phys, and (*C/D*) 6P+2Phys+GSR pipelines. In each subplot, the top 200 edges, ranked by *t*-values, are represented in brain space, and the percentage of connections that fall within and between brain networks identified by the Gordon parcellation are presented in matrix form immediately below. Anatomical renderings were generated using NeuroMarvl (http://immersive.erc.monash.edu.au/neuromarvl/). These plots can be viewed interactively at (*A*) http://bit.ly/2sXxZiU, (*B*) http://bit.ly/2rIwrWx, (C) http://bit.ly/2rYaua9, and (D) http://bit.ly/2t1VTdr.

Discussion

In this study, we compared the performance of sixteen different popular denoising strategies for rs-fMRI data using five different benchmarks and three datasets, and then used those benchmarks to inform interpretation of clinical group differences in functional connectivity. Our key results can be summarized as follow:

- Across most benchmarks, particularly QC-FC correlations, QC-FC distance-dependence, and HLM contrasts, the HMP class of pipelines performed well in low-motion data but poorly in high-motion data, indicating that simple regression of head motion parameters, even with expansion terms included, does not sufficiently correct the influence of motion on functional connectivity. Adding mean tissue physiological regressors to the HMP models improved denoising efficacy, but performance was still worse than other classes of pipelines.
- Compared to aCompCor, aCompCor50 pipelines showed lower QC-FC correlations but stronger QC-FC distance-dependence. They also showed the highest tDOF-loss of all pipelines. In general, the performance of aCompCor and aCompCor50 pipelines on most benchmarks was intermediate, with HMP pipelines showing poorer performance and ICA-AROMA pipelines showing better performance.
- ICA-AROMA pipelines were among the top performers across nearly all benchmarks. They showed good performance in relation to QC-FC correlations and QC-FC distance-dependence, identified no group differences in the HLM contrasts when results were corrected for multiple comparisons, and were associated with only moderate tDOF-loss. The variable tDOF-loss across participants resulted in significant

group differences between schizophrenia patients and HCs, which necessitated the inclusion of tDOF-loss as a covariate in group comparisons.

- In general, denoising pipelines incurring higher tDOF-loss were associated with lower TRT of functional connectivity estimates, such that the simplest 6HMP model showed the highest within- and between-session TRT and aCompCor50 pipelines showed the lowest. Given the relatively poor performance of the 6HMP model in denoising, it is possible that a substantial portion of reproducible BOLD signal is driven by head motion and other physiological confounds.
- Volume censoring yielded little additional benefit for the pipelines considered here. The primary benefit of censoring derived from the exclusion of participants with high levels of supra-threshold framewise displacements. Excluding such participants, particularly using the criteria employed in the spike regression pipeline used here, led to improvements in QC-FC benchmarks across HMP, aCompCor and ICA-AROMA pipelines.
- Including GSR generally improved QC-FC correlations but exacerbated QC-FC distance-dependence. It also appeared to shift the differences in the HLM contrasts to include more bidirectional effects (Fig. 6, $p<0.05$, uncorrected).
- In the CNP dataset, the direction of functional dysconnectivity, and the specific set of connections identified as being different, were contingent on the pipeline used. ICA-AROMA+2Phys was the most sensitive to group differences between HC and SCZ, which is encouraging given that it was also among the best performers across our benchmarks. It only identified differences in the HC>SCZ direction. Pipelines identifying differences in the SCZ>HC direction included either GSR or aCompCor.

Taken together, these findings indicate that, while no pipeline can successfully remove all confounding effects of head motion, ICA-AROMA is relatively successful, as assessed by nearly all our benchmarks. A particularly fruitful approach involves applying the method after exclusion of high motion participants based on typical spike regression criteria. The method incurs only a moderate loss of tDOF, and may boost sensitivity for detecting pathophysiologically-relevant differences between patients and controls. In the following sections, we discuss some of the key aspects of our results in more detail.

HMP pipelines

The most popular strategy used to correct for the confounding effects of head motion on BOLD signal variance involves simple linear regression using the canonical 6HMP, sometimes with mean tissue regressors and expansion terms included. Our evaluation indicates that this strategy is reasonable in low-motion datasets, as these methods showed low QC-FC correlations, moderate levels of QC-FC distance-dependence, and identified no differences in the HLM contrasts for the low-motion BMH sample. However, the HMP models performed the worst in the high-motion CNP data, suggesting that they lack the robustness to be applied generally to diverse datasets. These findings are consistent with past studies showing that the regression of head motion parameters is insufficient for removing the effect of motion on functional connectivity (Ciric et al., 2017; Yan et al., 2013a). Indeed, compared to aCompCor and ICA-AROMA pipelines, Ciric et al. (2017) also found that standard and expanded head motion parameter regression showed the highest QC-FC correlations. Yan et al. (2013a) also found that motion-BOLD correlations were still present across a wide range of expanded head motion models. Here, we found that HMP approaches were the most sensitive to differences between high- and low-motion CNP controls, with the 6HMP model identifying significant differences in functional connectivity for over half the network. Thus, analyses that rely on HMP models alone are likely to be heavily contaminated by motion unless overall levels of motion in the data are low.

aCompCor/aCompCor50 pipelines

Denoising pipelines incorporating variants of aCompCor are becoming increasingly popular, and have previously been shown to be more effective than mean WM/CSF regression for removing the relationship between motion and DVARS (Muschelli et al., 2014). Our evaluation indicates that these methods perform better than HMP models in high-motion data across most benchmarks, although they are often not as effective at denoising as pipelines using ICA-AROMA. The particular problem for the aCompCor/aCompCor50 pipelines is that they performed poorly in low-motion data, yielding significant QC-FC correlations and showing high levels of QC-FC distance-dependence. This dependence on levels of motion in the data suggests that, as with the HMP models, aCompCor/aCompCor50 pipelines are not sufficiently robust to be generally applicable to a diverse range of data. Furthermore, the aCompCor50 pipelines were also associated with very high levels of tDOF-loss, incorporating ~70-105 noise regressors, on average.

ICA-AROMA pipelines

Pipelines using ICA-AROMA performed consistently well across nearly all benchmarks. Performance was also consistent regardless of whether ICA-AROMA was combined with 2Phys, 8Phys, GSR or 4GSR regressors. This consistency is a desirable property of a robust noise correction approach. However, while Ciric et al. (2017) found that ICA-AROMAs performance was superior to that of HMP and mean WM/CSF models, it was outperformed by aCompCor on the QC-FC correlation benchmark. In fact, Ciric et al. found that ICA-AROMA showed FDR-corrected QC-FC correlations of up to 30%, roughly three times higher than our results for the CNP dataset. Indeed, our QC-FC correlation results showed that aCompCor/aCompCor50 only outperformed ICA-AROMA when GSR was included. While the exact reason for this discrepancy is unclear, it is unlikely to be due to Ciric et al.'s data containing high levels of motion because they used stringent exclusion criteria, removing participants with >0.2mm mean FD or with >20 volumes with FDs > 0.25mm. One possibility is that this discrepancy reflects dataset-to-dataset variability in denoising efficacy which, if true, further underscores the need for different studies to report QC-FC benchmarks (see *Recommendations* below).

Consistent with Ciric et al. (2017), we found that ICA-AROMA was the most effective at mitigating QC-FC distance-dependence. One caveat of ICA-AROMA is that it results in variable tDOF-loss across participants. As a result, tDOF-loss differed significantly between the SCZ and HC groups in the CNP dataset. We thus employed tDOF-loss as a nuisance covariate in our group analyses for all pipelines that included variable tDOF-loss.

Test-retest reliability

In general, pipelines that incurred lower tDOF-loss were associated with higher average TRT, with the simplest noise model – the 6HMP pipeline – showing the highest reliability. Given that this method was generally not successful in removing motion effects, this result suggests that a substantial fraction of reproducible BOLD signal may be driven by head motion and/or other physiological confounds. This conclusion is consistent with prior work indicating that head motion may be associated with a specific spatial pattern of activation, which may be represent a trait-like characteristic (Yan et al., 2013b; 2013a). Twin research has shown that head motion is indeed a heritable trait (Couvy-Duchesne et al., 2014). In particular, covariance between mean translational head motion, maximum translational motion, and mean rotation was related to a single latent head motion construct for

which half the variance was due to additive genetic factors. This result may partly explain why successful removal of motion effects reduces TRT.

Volume censoring

Volume censoring is another popular method for removing motion-related confounds from BOLD data. Explicitly removing high motion time points has been shown to be very effective at motion correction (Power et al., 2014; 2013; 2012). However, the approach can distort the temporal properties of the data, precluding analysis of spectral content or other time series properties (Sethi et al., 2017), and non-stationary dynamics (Hutchison et al., 2013; Zalesky et al., 2014).

In our analysis, the primary advantage of volume censoring came from the exclusion of participants with high levels of FDs, rather than the effect of censoring data in the remaining participants. As shown in Figure 4, excluding high-motion participants generally led to a dramatic reduction in QC-FC correlations, regardless of whether a HMP, aCompCor, or ICA-AROMA-based pipeline was used for denoising (except for ICA-AROMA+2Phys with scrubbing). There was only minor additional benefit of censoring after this exclusion. This result is consistent with Ciric et al. (2017), who showed small improvements to QC-FC correlations when applying spike regression to their 36P pipeline, which is equivalent to our 24HMP+8Phys+4GSR pipeline. Ciric et al. (2017) also found that scrubbing resulted in a slight increase of QC-FC correlations.

In our analysis, excluding participants using spike regression criteria resulted in more participants being excluded compared with using scrubbing. Accordingly, QC-FC correlations showed greater reductions for spike regression-based exclusion. Taken alone, the spike regression QC-FC correlation result implies that all denoising methods are as good each other once high motion participants have been excluded. However, neither exclusion nor censoring could reduce QC-FC distance-dependence for the HMP and aCompCor pipelines to levels matching ICA-AROMA+2Phys pre-censoring. This finding is also consistent with Ciric et al. (2017), who found that QC-FC distance-dependence was higher for 24HMP+8Phys+4GSR *with* censoring compared to ICA-AROMA+2Phys *without* censoring.

Our results thus suggest that a very important first step is to exclude participants who have high levels of censored data. In particular, the criteria applied as part of the spike regression pipeline implemented here, in which people were excluded if the removal of spikes, defined as instances where $FD_{Jenk} > 0.25$mm, results in less than 4-minutes of remaining data, were particularly effective in mitigating the effects of motion.

Global signal regression

GSR is one of the most controversial preprocessing steps in the fMRI literature. It has been shown to improve correction for motion-related artefacts and other physiological confounds (Power et al., 2017) and it improves the spatial specificity and anatomical plausibility of seed-based functional connectivity maps (Fox et al., 2009). However, GSR mathematically forces the distribution of correlations between voxels to be centred around zero, introducing anti-correlations that may be artefactual (Fox et al., 2009).

In our analysis, including GSR led to small improvements in QC-FC correlations for HMP pipelines, major improvements for aCompCor pipelines, and a slight decrement in performance for ICA-AROMA pipelines. It was also often associated with a higher QC-FC distance-dependence. Most importantly however, group differences in the CNP dataset only emerged for the SCZ>HC contrast when either GSR or aCompCor was used. This result may reflect the transforming effect that GSR can have on between-group contrasts of functional connectivity estimates (Saad et al., 2012). The fact that SCZ>HC differences also emerged with aCompCor may be related to the extent to which the noise components in the aCompCor model (particularly the first PC) correlate with the global signal.

The inherent assumption of GSR is that the global signal is mostly dominated by the noise rather than the true neural signal. The problem is that the global signal samples from regions of grey matter. These grey matter signals will contribute to the global signal to the extent that they are coupled with each other. In other words, tightly coupled sets of grey matter voxels (i.e., putative functional networks of brain), will contribute strongly to the global signal. Removal of the global signal will thus subtract the common components of these networks from other parts of the brain, potentially introducing anti-correlations between brain regions that otherwise show weak functional connectivity. The challenge is that it is difficult to predict this affect *a priori* – the impact of GSR will depend on the initial covariance structure of grey matter (and WM/CSF) signals. Consistent with this view, Saad et al. (2012) and Gotts et al. (2013) have used both modelling and experiments to show that the impact of GSR is greatest when two groups differ in the underlying correlation structure of the network. For instance, if the size of one network differs between patients and HCs, the global signal will contain a greater contribution from this network in one group relative the other, causing GSR to affect this system differently between the two groups, and thus potentially resulting in spurious group

differences. Critically, Saad et al. (2012) derived an equation for determining the correlation structure of a network post-GSR given its architecture pre-GSR. The model of GSR transformation was later shown to explain >95% of the variance in empirical functional connectivity measures post-GSR (Gotts et al., 2013), demonstrating the explanatory power of this basic mathematical property of GSR.

Nonetheless, failing to remove the global signal may also conceivably result in spurious group differences, particularly when two groups differ in the amount of global noise. For example, Power et al. (2017) have recently shown that GSR is the only effective method for removing large-scale, global signal fluctuations in fMRI time series that are mainly attributable to respiration. The efficacy of GSR stood in stark contrast to model-based correction methods that used physiological recordings, which were only partially effective in removing these global fluctuations. If two groups differ in respiratory or other physiological processes, which is plausible in patient cohorts who may have heightened anxiety in the scanner environment, then these differences may impact variations in global noise. It is an empirical question as to whether these effects are sufficiently severe to justify the use of GSR, given its capacity to distort the group differences in the correlation structure of the data.

Sensitivity to clinical differences

The analysis of the BMH dataset revealed no differences between OCD patients and HCs for any pipeline, whereas analysis of the CNP dataset identified differences between SCZ and HCs for several denoising approaches. One interpretation of this discrepancy is that the group differences in the CNP dataset may simply reflect the higher levels of motion in this dataset. However, differences between SCZ and HC were identified even using pipelines such as ICA-AROMA+2Phys, that were relatively successful in removing motion-related effects according to our benchmarks. Thus, it is possible that disease-related brain changes are more pronounced in schizophrenia than OCD, at least as measured using the current analysis strategy (i.e., brain-wide comparison of pair-wise functional connectivity in a 333-node or 264-node network).

A reassuring finding is that ICA-AROMA+2Phys, which was the generally one of the best performers across all denoising benchmarks, was the most sensitive to differences between SCZ and HC, identifying the largest sub-network of disease-related changes compared to the other pipelines. The performance of this method on the various other benchmarks studied here increases confidence that these changes are indeed disorder-related. These two properties –

successful performance on denoising benchmarks and enhanced sensitivity to putative group differences – are fundamental to an effective preprocessing pipeline.

A more troubling result is that the specific subset of connections showing differences, and even the direction of those differences, is highly dependent on the preprocessing pipeline used. For example, if we use the 24HMP+8Phys+4GSR, which remains a commonly used approach in the literature, we conclude that schizophrenia is associated with increased functional connectivity within and between visual and motor networks (Fig. 10A). If we use ICA-AROMA+2Phys, we conclude that schizophrenia is associated with reduced connectivity between default-mode and executive networks (Fig. 10B). Such variability, caused by differences in data processing strategies, will result in a literature littered with inconsistent findings, leading to difficulties in replication and a major impediment to the development of valid models of disease pathophysiology. It is thus imperative to accurately characterize the effects of motion in any group comparison. To this end, we present some simple strategies in the following section.

Recommendations

Figure 11 presents a suggested workflow that can help mitigate and characterize the effects of motion on functional connectivity, and which will ensure that sufficient data are reported to allow readers to make their own judgements regarding the success of a particular denoising procedure. The basic workflow involves the following steps. First, inspection of each participant's processed EPI data as a time series (i.e., "carpet plot") alongside quality control metrics such as FD and DVARS (Power, 2016). We provide code for this analysis (https://github.com/lindenmp/rs-fMRI) but also note that the recently available toolbox, MRIQC (https://github.com/poldracklab/mriqc), provides extensive documentation on this practice. Second, removal of participants with a high proportion of contaminated volumes. Specifically, we recommend discarding participants who have < 4-minutes of uncontaminated EPI data, where contaminated data is defined as any volume with >0.25mm FD_{Jenk}. Third, pre-processing and denoising. We found that ICA-AROMA+2Phys (Pruim et al., 2015b) performs well at mitigating the effects of motion on functional connectivity and was the most sensitive to group differences between patients with schizophrenia and healthy controls. The final step is to report quality control benchmarks. Specifically, QC-FC correlations, QC-FC distance-dependence, group differences in functional connectivity between high- and low-motion healthy controls (where sample size permits) and tDOF-loss (where appropriate). These benchmarks will provide critical insights to the effectiveness of data denoising in a given study,

providing readers with a more complete understanding of the strengths and weaknesses of a given analysis.

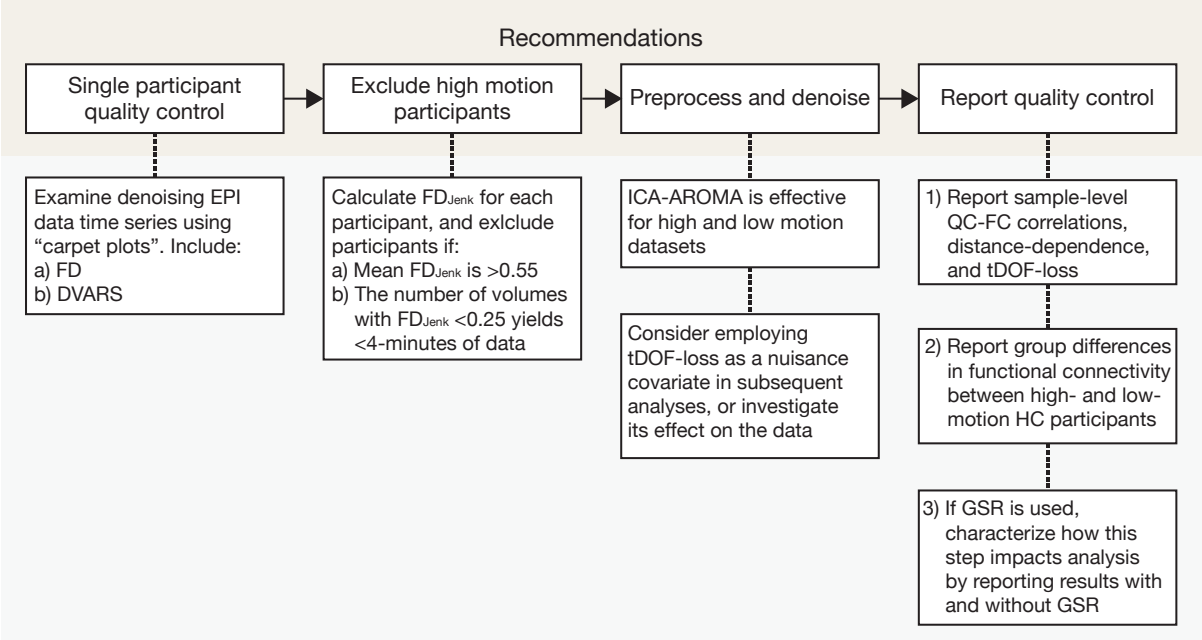

Figure 11. Recommended workflow for characterising and reporting motion characteristics in resting-state fMRI analyses.

Limitations

The lack of a ground truth makes benchmarking the effectiveness of any noise correction strategy difficult. We compensated for this limitation by conducting a comprehensive investigation of noise correction pipelines, looking for convergence between QC-FC benchmark analyses, loss in temporal degrees of freedom, within- and between-session reliability, and sensitivity to clinical group differences in functional connectivity. Nonetheless, we cannot distinguish true positive/negatives from false positives/negatives in case-control analyses with absolute certainty.

We focused here on simple denoising procedures that are readily available and which can be applied to any dataset. Alternative, prospective methods for motion correction that we did not consider include acquiring multi-echo data (Kundu et al., 2013) and actual monitoring of head motion in the scanner (Herbst et al., 2013; Maclaren et al., 2012). These techniques have shown great promise and provide a fruitful way of addressing the problem of head motion in future studies.

Other methods that can be applied to any data, and which we have not considered here, are ICA-FIX (Griffanti et al., 2014; Salimi-Khorshidi et al., 2014) and wavelet despiking (Patel et al., 2014). A limitation of ICA-FIX is that it requires curation of a training dataset with the same acquisition protocol as the analysis sample. Therefore, it requires relatively large samples so that a subset of individuals can be held out as the training set. Wavelet despiking does not suffer from this limitation, and has been shown to be more effective than volume censoring in

removing the effects of head motion (Patel et al., 2014). However, the technique does not explicitly incorporate a model of signal noise and requires the setting of a free parameter that must be adapted to different kinds of data. We focused here on relatively automated methods that require little user input.

Finally, we did not examine the efficacy of these denoising pipelines for cleaning multiband fMRI data. Such data can have a distinct noise structure, which can be characterised with ICA as components with sparse and evenly spaced slices (Griffanti et al., 2016), but which are not likely to be identified as noise by the restricted feature set used in ICA-AROMA. Thus, for multiband data, techniques such as ICA-FIX, in which such noise components can be explicitly identified by users, may be more appropriate.

Conclusions

In summary, we comprehensively examined a wide variety of commonly adopted noise correction methods for resting state fMRI data. We found that ICA-AROMA (Pruim et al., 2015b; 2015a) performed best across a range of benchmarks and was most sensitive to functional dysconnectivity in clinical groups. Our work contributes to the growing body of literature (Ciric et al., 2017; Power et al., 2015; Satterthwaite et al., 2013; Yan et al., 2013a) highlighting the suboptimal performance of many common noise correction methods used in the literature. Crucially, our results underscore the importance of reporting the residual relationship between in-scanner movement and functional connectivity alongside case-control comparisons in clinical neuroimaging studies. Reporting QC-FC benchmarks will assist investigators and readers to understand which group differences are more likely to be true positives and which may be spurious.

References

Avants, B., Epstein, C., Grossman, M., Gee, J., 2008. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12, 26–41. doi:10.1016/j.media.2007.06.004

Beckmann, C.F., DeLuca, M., Devlin, J.T., Smith, S.M., 2005. Investigations into resting-state connectivity using independent component analysis. Philosophical Transactions of the Royal Society B: Biological Sciences 360, 1001–1013. doi:10.1098/rstb.2005.1634

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage 37, 90–101. doi:10.1016/j.neuroimage.2007.04.042

Birn, R.M., 2012. The role of physiological noise in resting-state functional connectivity. NeuroImage 62, 864–870. doi:10.1016/j.neuroimage.2012.01.016

Birn, R.M., Cornejo, M.D., Molloy, E.K., Patriat, R., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V., 2014. The Influence of Physiological Noise Correction on Test–Retest Reliability of Resting-State Functional Connectivity. Brain Connectivity 4, 511–522. doi:10.1089/brain.2014.0284

Chen, G., Chen, G., Xie, C., Ward, B.D., Li, W., Antuono, P., Li, S.-J., 2012. A method to determine the necessity for global signal regression in resting-state fMRI studies. Magn Reson Med 68, 1828–1835. doi:10.1002/mrm.24201

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. NeuroImage 1–22. doi:10.1016/j.neuroimage.2017.03.020

Cole, M.W., Bassett, D.S., Power, J.D., Braver, T.S., Petersen, S.E., 2014. Intrinsic and Task-Evoked Network Architectures of the Human Brain. Neuron 83, 238–251. doi:10.1016/j.neuron.2014.05.014

Couvy-Duchesne, B., Blokland, G.A.M., Hickie, I.B., Thompson, P.M., Martin, N.G., de Zubicaray, G.I., McMahon, K.L., Wright, M.J., 2014. Heritability of head motion during resting state functional MRI in 462 healthy twins. NeuroImage 102, 424–434. doi:10.1016/j.neuroimage.2014.08.010

Dandash, O., Ben J Harrison, Adapa, R., Gaillard, R., Giorlando, F., Wood, S.J., Fletcher,

P.C., Fornito, A., 2014. Selective Augmentation of Striatal Functional Connectivity Following NMDA Receptor Antagonism: Implications for Psychosis. Neuropsychopharmacology 40, 622–631. doi:10.1038/npp.2014.210

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. Nature Publishing Group 18, 1664–1671. doi:10.1038/nn.4135

Fornito, A., Bullmore, E.T., 2010. What can spontaneous fluctuations of the blood oxygenation-level-dependent signal tell us about psychiatric disorders? Current Opinion in Psychiatry 23, 239–249. doi:10.1097/YCO.0b013e328337d78d

Fornito, A., Harrison, B.J., Goodby, E., Dean, A., Ooi, C., Nathan, P.J., Lennox, B.R., Jones, P.B., Suckling, J., Bullmore, E.T., 2013. Functional Dysconnectivity of Corticostriatal Circuitry as a Risk Phenotype for Psychosis. JAMA Psychiatry 70, 1143–9. doi:10.1001/jamapsychiatry.2013.1976

Fornito, A., Yoon, J., Zalesky, A., Bullmore, E.T., Carter, C.S., 2011a. General and Specific Functional Connectivity Disturbances in First-Episode Schizophrenia During Cognitive Control Performance. BPS 70, 64–72. doi:10.1016/j.biopsych.2011.02.019

Fornito, A., Zalesky, A., Bassett, D.S., Meunier, D., Ellison-Wright, I., Yücel, M., Wood, S.J., Shaw, K., O'Connor, J., Nertney, D., Mowry, B.J., Pantelis, C., Bullmore, E.T., 2011b. Genetic Influences on Cost-Efficient Organization of Human Cortical Functional Networks. Journal of Neuroscience 31, 3261–3270. doi:10.1523/JNEUROSCI.4858-10.2011

Fornito, A., Zalesky, A., Bullmore, E., 2016. Fundamentals of Human Imaging Connectomics. Elsevier.

Fornito, A., Zalesky, A., Pantelis, C., Bullmore, E.T., 2012. Schizophrenia, neuroimaging and connectomics. NeuroImage 62, 2296–2314. doi:10.1016/j.neuroimage.2011.12.090

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8, 700–711. doi:10.1038/nrn2201

Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E., 2007. Intrinsic Fluctuations within Cortical Systems Account for Intertrial Variability in Human Behavior. Neuron 56, 171–184. doi:10.1016/j.neuron.2007.08.023

Fox, M.D., Zhang, D., Snyder, A.Z., Raichle, M.E., 2009. The Global Signal and Observed Anticorrelated Resting State Brain Networks. Journal of Neurophysiology 101, 3270–

3283. doi:10.1152/jn.90777.2008

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J., Turner, R., 1996. Movement-Related effects in fMRI time-series. Magn Reson Med 35, 346–355. doi:10.1002/mrm.1910350312

Glahn, D.C., Winkler, A.M., Kochunov, P., Almasy, L., Duggirala, R., Carless, M.A., Curran, J.C., Olvera, R.L., Laird, A.R., Smith, S.M., Beckmann, C.F., Fox, P.T., Blangero, J., 2010. Genetic control over the resting brain. Proceedings of the National Academy of Sciences 107, 1223–1228. doi:10.1073/pnas.0909969107

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. Cereb. Cortex 26, 288–303. doi:10.1093/cercor/bhu239

Gotts, S.J., Saad, Z.S., Jo, H.J., Wallace, G.L., Cox, R.W., Martin, A., 2013. The perils of global signal regression for group comparisons: a case study of Autism Spectrum Disorders. Front. Hum. Neurosci. 7, 1–21. doi:10.3389/fnhum.2013.00356

Griffanti, L., Douaud, G., Bijsterbosh, J., Evangelisti, S., Alfaro-Almagro, F., Glasser, M.F., Duff, E.P., Fitzgibbon, S., Westphal, R., Carone, D., Beckmann, C.F., Smith, S.M., 2016. Hand classification of fMRI ICA noise components. NeuroImage 1–18. doi:10.1016/j.neuroimage.2016.12.036

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. NeuroImage 95, 232–247. doi:10.1016/j.neuroimage.2014.03.034

Herbst, M., Maclaren, J., Lovell-Smith, C., Sostheim, R., Egger, K., Harloff, A., Korvink, J., Hennig, J., Zaitsev, M., 2013. Reproduction of motion artifacts for performance analysis of prospective motion correction in MRI. Magn Reson Med 71, 182–190. doi:10.1002/mrm.24645

Hutchison, R.M., Womelsdorf, T., Allen, E.A., Bandettini, P.A., Calhoun, V.D., Corbetta, M., Penna, Della, S., Duyn, J.H., Glover, G.H., Gonzalez-Castillo, J., Handwerker, D.A., Keilholz, S., Kiviniemi, V., Leopold, D.A., de Pasquale, F., Sporns, O., Walter, M., Chang, C., 2013. Dynamic functional connectivity: Promise, issues, and interpretations. NeuroImage 80, 360–378. doi:10.1016/j.neuroimage.2013.05.079

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, 825–841. doi:10.1006/nimg.2002.1132

Kundu, P., Brenowitz, N.D., Voon, V., Worbe, Y., Vertes, P.E., Inati, S.J., Saad, Z.S., Bandettini, P.A., Bullmore, E.T., 2013. Integrated strategy for improving functional connectivity mapping using multiecho fMRI. Proceedings of the National Academy of Sciences 110, 16187–16192. doi:10.1073/pnas.1301725110

Lemieux, L., Salek-Haddadi, A., Lund, T.E., Laufs, H., Carmichael, D., 2007. Modelling large motion events in fMRI studies of patients with epilepsy. Magnetic Resonance Imaging 25, 894–901. doi:10.1016/j.mri.2007.03.009

Maclaren, J., Armstrong, B.S.R., Barrows, R.T., Danishad, K.A., Ernst, T., Foster, C.L., Gumus, K., Herbst, M., Kadashevich, I.Y., Kusik, T.P., Li, Q., Lovell-Smith, C., Prieto, T., Schulze, P., Speck, O., Stucht, D., Zaitsev, M., 2012. Measurement and Correction of Microscopic Head Motion during Magnetic Resonance Imaging of the Brain. PLoS ONE 7, e48088–9. doi:10.1371/journal.pone.0048088

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? NeuroImage 44, 893–905. doi:10.1016/j.neuroimage.2008.09.036

Muschelli, J., Nebel, M.B., Caffo, B.S., Barber, A.D., Pekar, J.J., Mostofsky, S.H., 2014. Reduction of motion-related artifacts in resting state fMRI using aCompCor. NeuroImage 96, 22–35. doi:10.1016/j.neuroimage.2014.03.028

Patel, A.X., Kundu, P., Rubinov, M., Jones, P.S., Vértes, P.E., Ersche, K.D., Suckling, J., Bullmore, E.T., 2014. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. NeuroImage 95, 287–304. doi:10.1016/j.neuroimage.2014.03.012

Pettersson-Yeo, W., Allen, P., Benetti, S., McGuire, P., Mechelli, A., 2011. Dysconnectivity in schizophrenia: Where are we now? Neuroscience and Biobehavioral Reviews 35, 1110–1124. doi:10.1016/j.neubiorev.2010.11.004

Poldrack, R.A., Congdon, E., Triplett, W., Gorgolewski, K.J., Karlsgodt, K.H., Mumford, J.A., Sabb, F.W., Freimer, N.B., London, E.D., Cannon, T.D., Bilder, R.M., 2016. A phenome-wide examination of neural and cognitive function. Sci. Data 3, 160110–12. doi:10.1038/sdata.2016.110

Power, J.D., 2016. A simple but useful way to assess fMRI scan qualities. NeuroImage 1–9. doi:10.1016/j.neuroimage.2016.08.009

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2013. Steps toward

optimizing motion artifact removal in functional connectivity MRI; a reply to Carp. NeuroImage 76, 439–441. doi:10.1016/j.neuroimage.2012.03.017

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, 2142–2154. doi:10.1016/j.neuroimage.2011.10.018

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2011. Functional Network Organization of the Human Brain. Neuron 72, 665–678. doi:10.1016/j.neuron.2011.09.006

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. doi:10.1016/j.neuroimage.2013.08.048

Power, J.D., Plitt, M., Laumann, T.O., Martin, A., 2017. Sources and implications of whole-brain fMRI signals in humans. NeuroImage 146, 609–625. doi:10.1016/j.neuroimage.2016.09.038

Power, J.D., Schlaggar, B.L., Petersen, S.E., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. NeuroImage 105, 536–551. doi:10.1016/j.neuroimage.2014.10.044

Pruim, R.H.R., Mennes, M., Buitelaar, J.K., Beckmann, C.F., 2015a. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. NeuroImage 112, 278–287. doi:10.1016/j.neuroimage.2015.02.063

Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015b. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage 112, 267–277. doi:10.1016/j.neuroimage.2015.02.064

Saad, Z.S., Gotts, S.J., Murphy, K., Chen, G., Jo, H.J., Martin, A., Cox, R.W., 2012. Trouble at Rest: How Correlation Patterns and Group Differences Become Distorted After Global Signal Regression. Brain Connectivity 2, 25–32. doi:10.1089/brain.2012.0080

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. NeuroImage 90, 449–468. doi:10.1016/j.neuroimage.2013.11.046

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the

preprocessing of resting-state functional connectivity data. NeuroImage 64, 240–256. doi:10.1016/j.neuroimage.2012.08.052

Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakonarson, H., Gur, R.C., Gur, R.E., 2012. Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. NeuroImage 60, 623–632. doi:10.1016/j.neuroimage.2011.12.063

Sethi, S.S., Zerbi, V., Wenderoth, N., Fornito, A., Fulcher, B.D., 2017. Structural connectome topology relates to regional BOLD signal dynamics in the mouse brain. Chaos 27, 047405–15. doi:10.1063/1.4979281

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86, 420–428. doi:10.1037/0033-2909.86.2.420

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proceedings of the National Academy of Sciences 106, 13040–13045. doi:10.1073/pnas.0905267106

Van Dijk, K.R.A., Sabuncu, M.R., Buckner, R.L., 2012. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage 59, 431–438. doi:10.1016/j.neuroimage.2011.07.044

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R.C., Di Martino, A., Li, Q., Zuo, X.-N., Castellanos, F.X., Milham, M.P., 2013a. A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. NeuroImage 76, 183–201. doi:10.1016/j.neuroimage.2013.03.004

Yan, C.-G., Craddock, R.C., Milham, M.P., He, Y., 2013b. Addressing head motion dependencies for small-world topologies in functional connectomics 1–19. doi:10.3389/fnhum.2013.00910/abstract

Zalesky, A., Fornito, A., Bullmore, E.T., 2010. Network-based statistic: Identifying differences in brain networks. NeuroImage 53, 1197–1207. doi:10.1016/j.neuroimage.2010.06.041

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L., Breakspear, M., 2014. Time-resolved resting-state brain networks. Proceedings of the National Academy of Sciences 111, 10341–10346. doi:10.1073/pnas.1400181111

Zalesky, A., Fornito, A., Seal, M.L., Cocchi, L., Westin, C.F., Bullmore, E.T., Egan, G.F., Pantelis, C., 2011. Disrupted Axonal Fiber Connectivity in Schizophrenia. BPS 69, 80–89. doi:10.1016/j.biopsych.2010.08.022

Zuo, X.-N., Anderson, J.S., Bellec, P., Birn, R.M., Biswal, B.B., Blautzik, J., Breitner, J.C.S., Buckner, R.L., Calhoun, V.D., Castellanos, F.X., Chen, A., Chen, B., Chen, J., Chen, X., Colcombe, S.J., Courtney, W., Craddock, R.C., Di Martino, A., Dong, H.-M., Fu, X., Gong, Q., Gorgolewski, K.J., Han, Y., He, Y., He, Y., Ho, E., Holmes, A., Hou, X.-H., Huckins, J., Jiang, T., Jiang, Y., Kelley, W., Kelly, C., King, M., LaConte, S.M., Lainhart, J.E., Lei, X., Li, H.-J., Li, K., Li, K., Lin, Q., Liu, D., Liu, J., Liu, X., Liu, Y., Lu, G., Lu, J., Luna, B., Luo, J., Lurie, D., Mao, Y., Margulies, D.S., Mayer, A.R., Meindl, T., Meyerand, M.E., Nan, W., Nielsen, J.A., O Connor, D., Paulsen, D., Prabhakaran, V., Qi, Z., Qiu, J., Shao, C., Shehzad, Z., Tang, W., Villringer, A., Wang, H., Wang, K., Wei, D., Wei, G.-X., Weng, X.-C., Wu, X., Xu, T., Yang, N., Yang, Z., Zang, Y.-F., Zhang, L., Zhang, Q., Zhang, Z., Zhang, Z., Zhao, K., Zhen, Z., Zhou, Y., Zhu, X.-T., Milham, M.P., 2014. An open science resource for establishing reliability and reproducibility in functional connectomics. Sci. Data 1, 140049–13. doi:10.1038/sdata.2014.49

Supplementary Figures

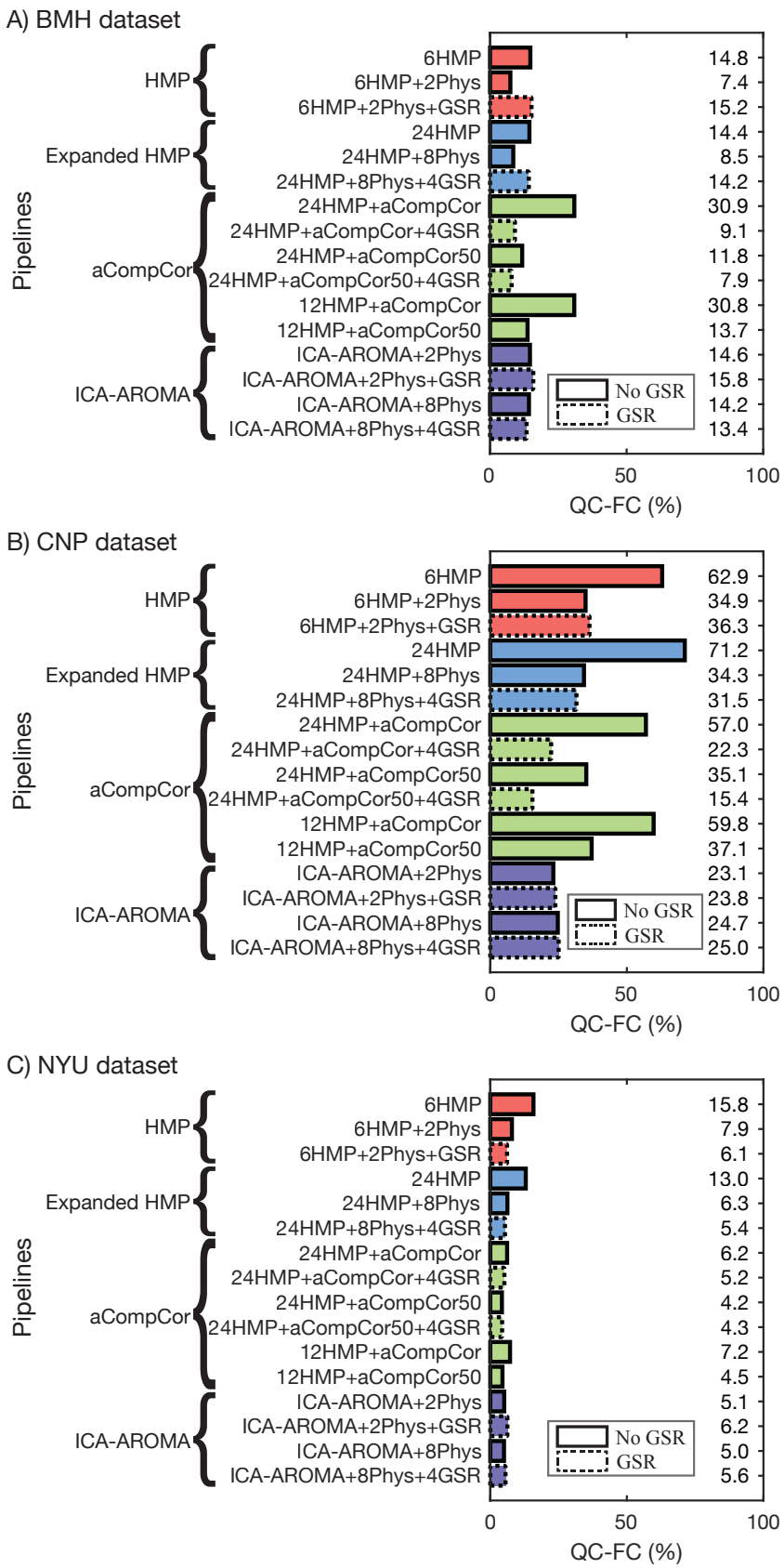

Figure S1. The residual effect of in-scanner motion on functional connectivity after noise correction with one of sixteen different rfMRI pre-processing pipelines. Functional connectivity at each edge was correlated with a summary metric of in-scanner movement across the entire sample (QC-FC correlations) for three separate datasets. The proportion of

functional connections that correlated significantly (p<0.05, **uncorrected**) with subject motion are shown for each pipeline for the BMH (*A*), CNP (*B*), and NYU (*C*) datasets separately.

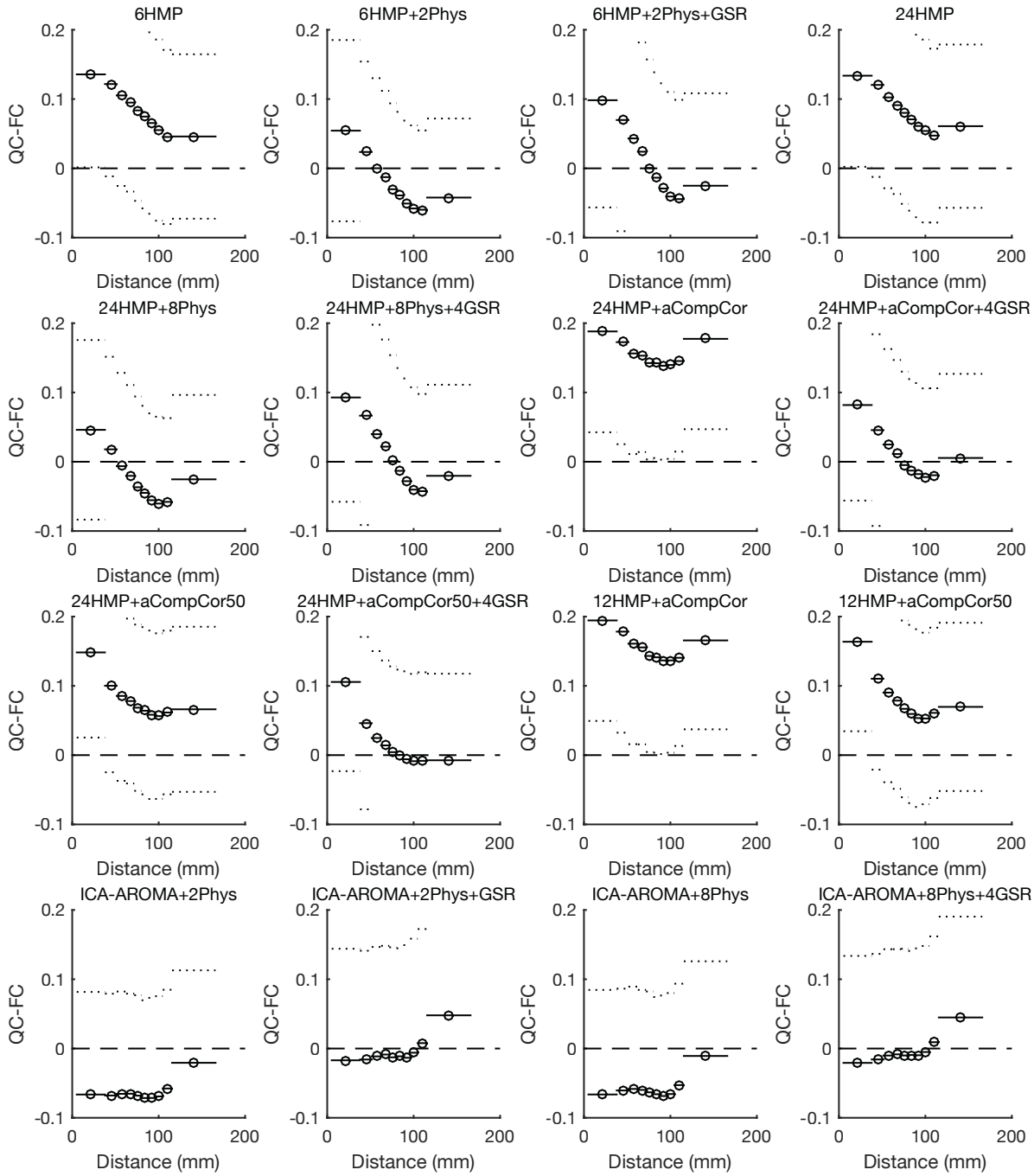

Figure S2. QC-FC distance dependence for each pipeline applied to the BMH dataset.

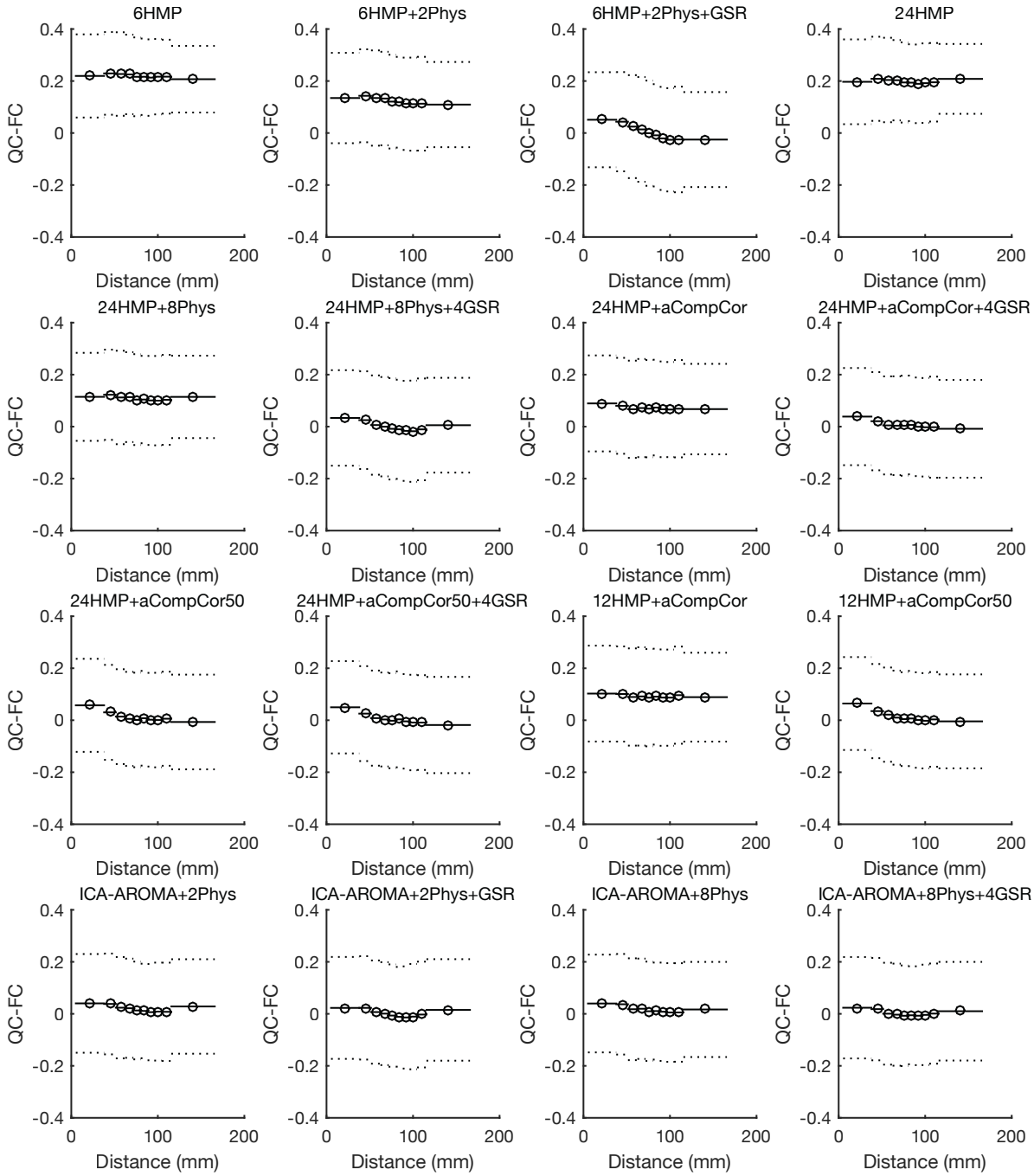

Figure S3. QC-FC distance dependence for each pipeline applied to the NYU dataset.

Primary analyses with the Power parcellation instead of the Gordon parcellation

The following figures mirror Figures 1, 2, 4, 6, 7 and 8 found in the main text but use the Power parcellation instead of the Gordon parcellation.

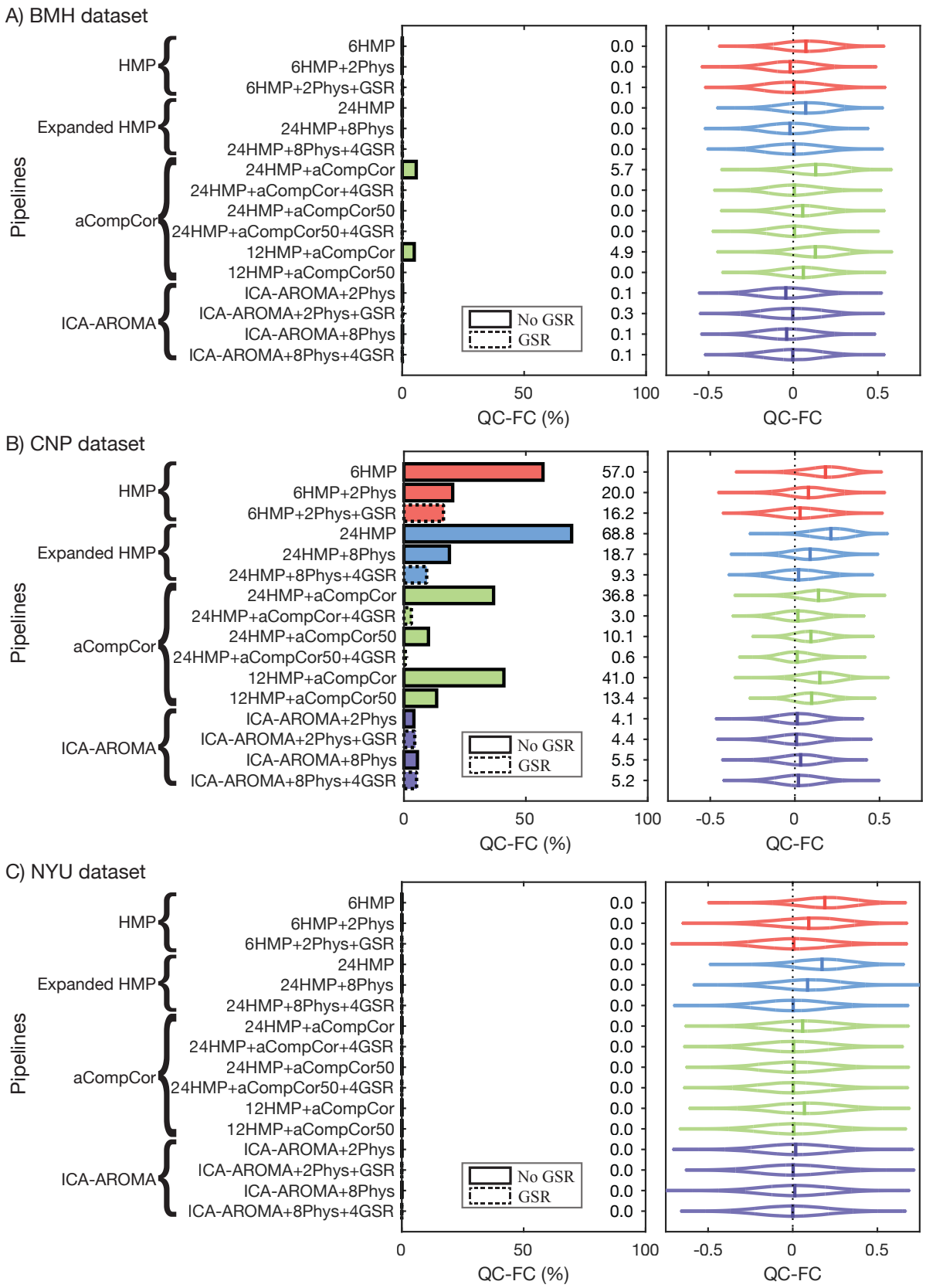

Figure S4. QC-FC correlations as per Figure 1 from main text but using the Power parcellation instead of the Gordon parcellation.

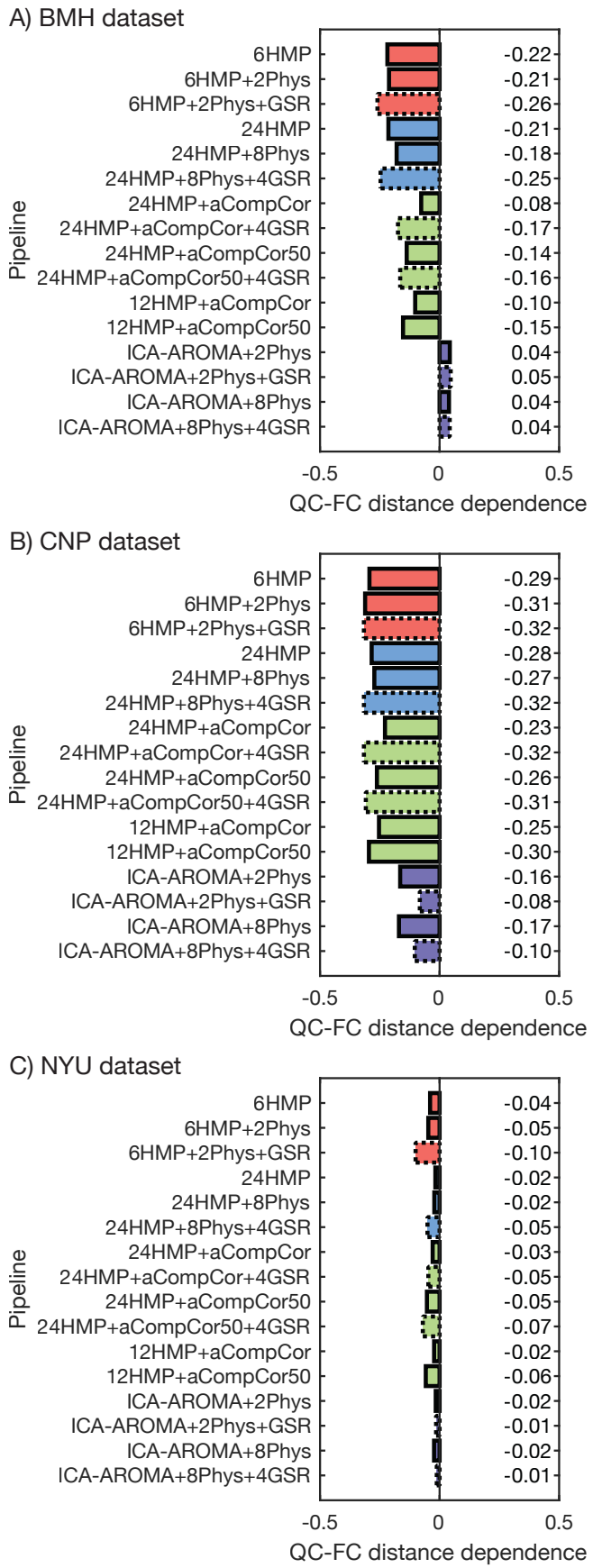

Figure S5. QC-FC distance dependence as per Figure 2 from main text but using the Power parcellation instead of the Gordon parcellation.

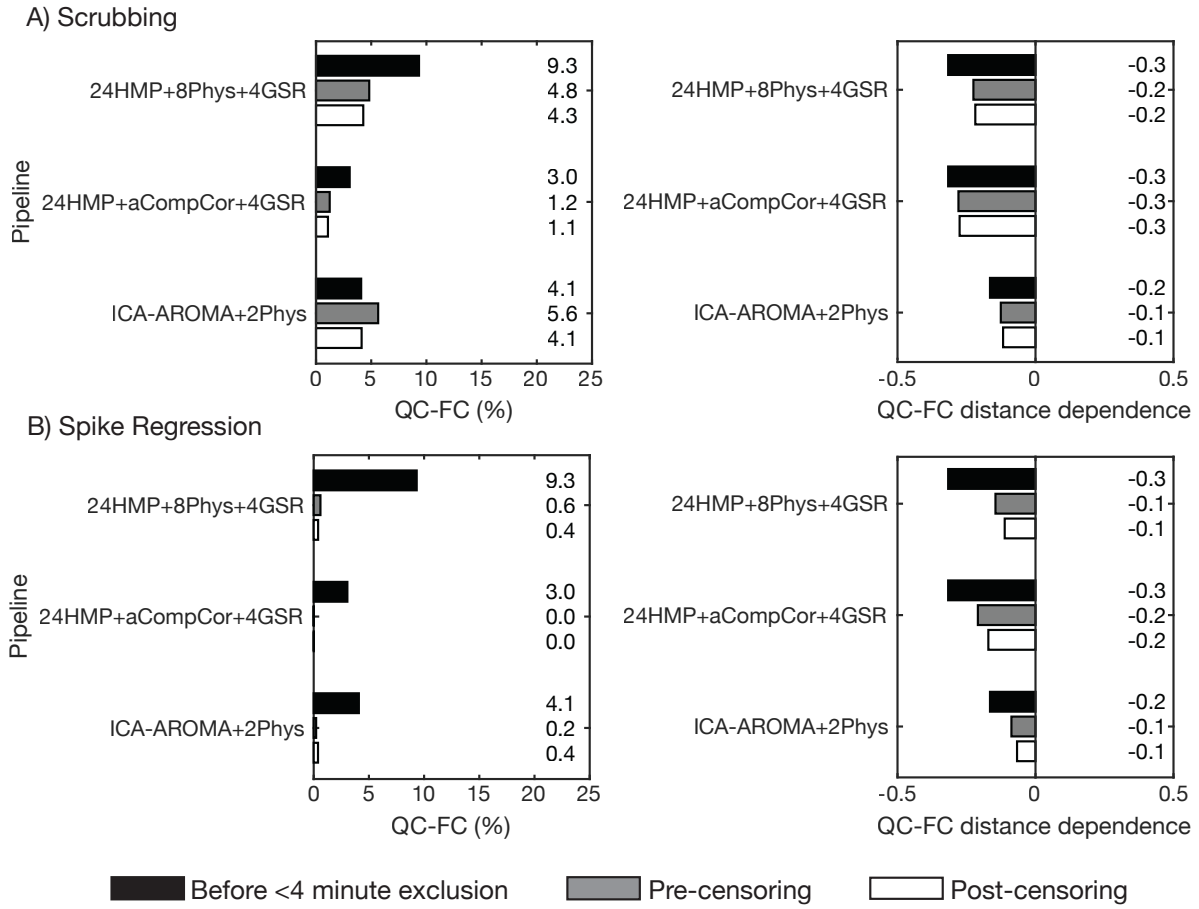

Figure S6. Volume censoring as per Figure 4 from main text but using the Power parcellation instead of the Gordon parcellation.

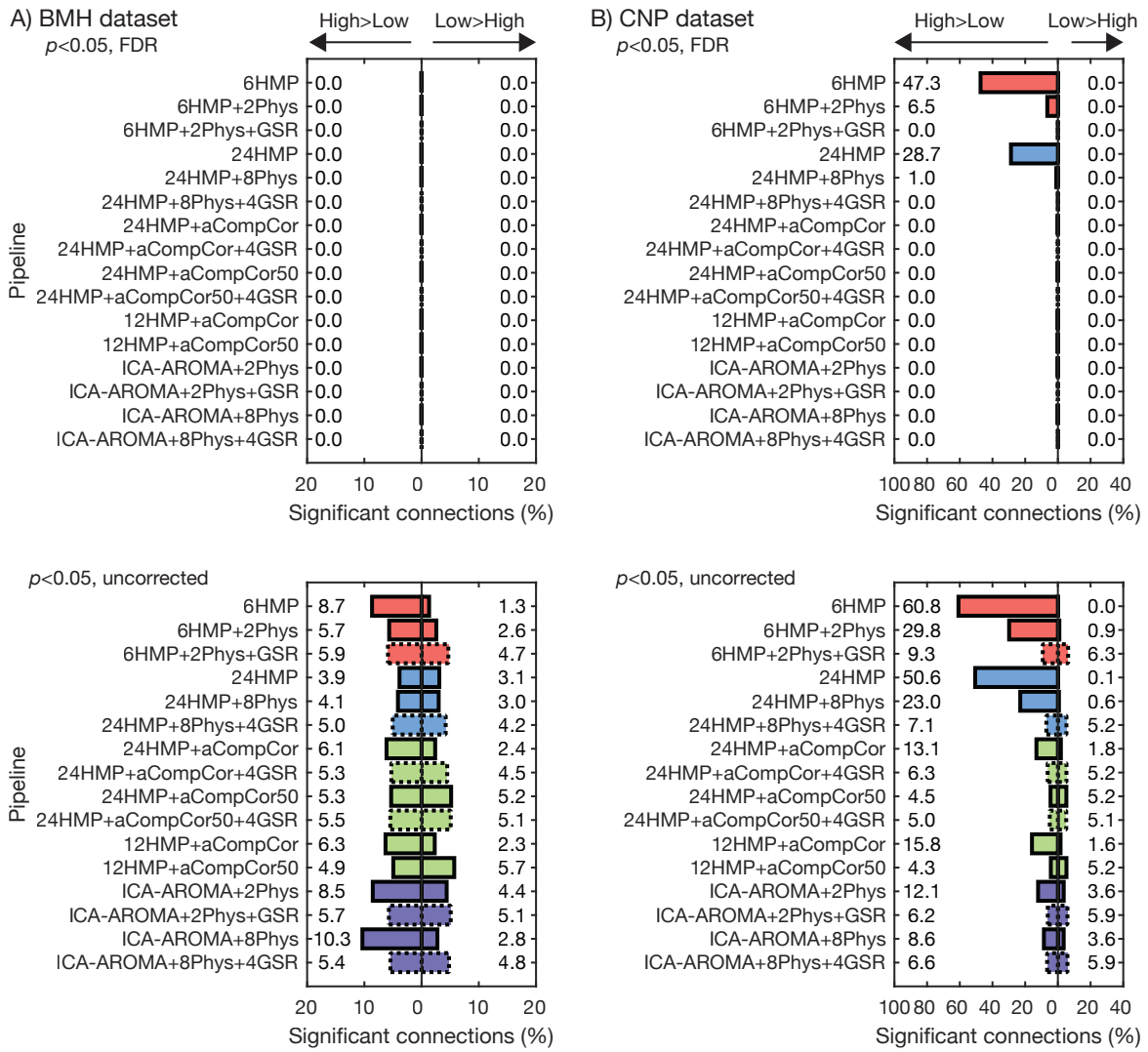

Figure S7. High/low-motion contrasts as per Figure 6 from main text but using the Power parcellation instead of the Gordon parcellation.

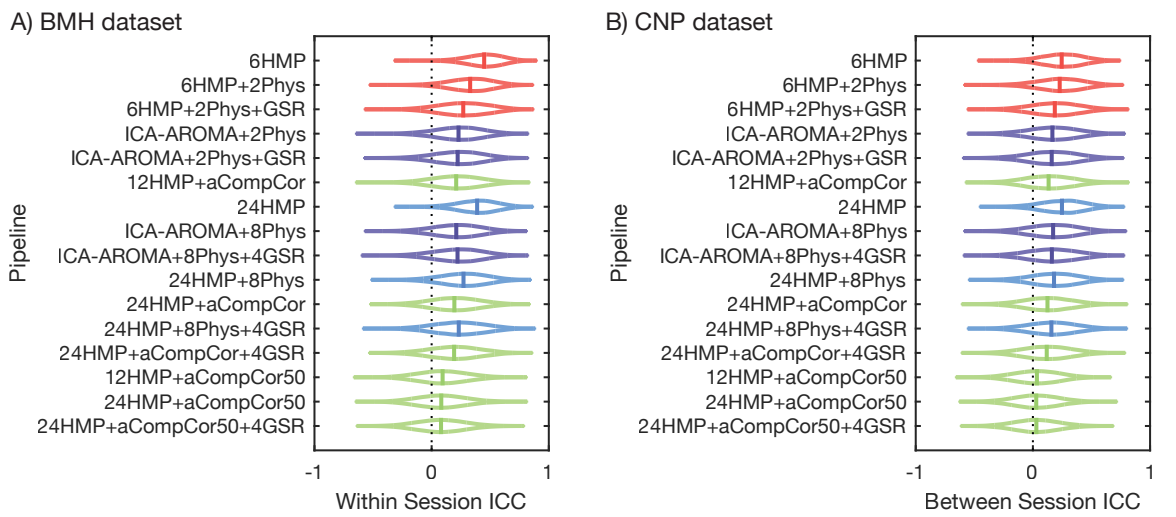

Figure S8. TRT as per Figure 7 from main text but using the Power parcellation instead of the Gordon parcellation.

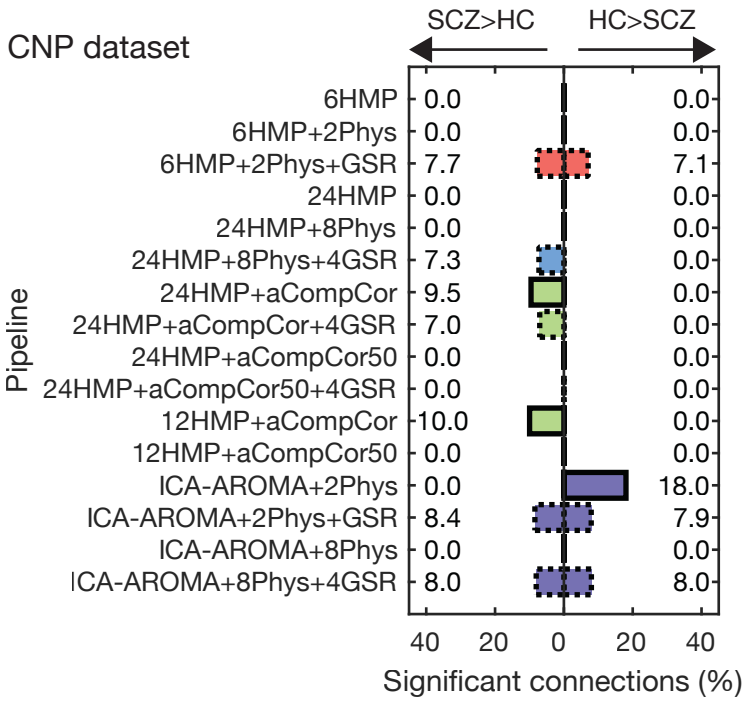

Figure S8. Case-control contrasts as per Figure 8 from main text but using the Power parcellation instead of the Gordon parcellation.