White Matter Denoising Improves the Identifiability of Large-Scale Networks and Reduces the Effects of Motion in fMRI Functional Connectivity

Michalis Kassinopoulos[1], Georgios D. Mitsis[2]

[1]Graduate Program in Biological and Biomedical Engineering, McGill University, Montreal, QC, Canada

[2]Department of Bioengineering, McGill University, Montreal, QC, Canada

Abstract

It is well established that confounding factors related to head motion and physiological processes (e.g. cardiac and breathing activity) should be taken into consideration when analyzing and interpreting results in fMRI studies. However, even though recent studies aimed to evaluate the performance of different preprocessing pipelines there is still no consensus on the *optimal* strategy. This may be partly because the quality control (QC) metrics used to evaluate differences in performance across pipelines often yielded contradictory results. Importantly, noise correction techniques based on physiological recordings or expansions of tissue-based techniques such as aCompCor have not received enough attention. Here, to address the aforementioned issues, we evaluate the performance of a large range of pipelines by using previously proposed and novel quality control (QC) metrics. Specifically, we examine the effect of three commonly used practices: 1) Removal of nuisance regressors from fMRI data, 2) discarding motion-contaminated volumes (i.e., scrubbing) before regression, and 3) low-pass filtering the data and the nuisance regressors before their removal. To this end, we propose a framework that summarizes the scores from eight QC metrics to a reduced set of two QC metrics that reflect the signal-to-noise ratio (SNR) and the reduction in motion artifacts and biases in the preprocessed fMRI data. Using resting-state fMRI data from the Human Connectome Project, we show that the best data quality, is achieved when the global signal (GS) and about 17% of principal components from white matter (WM) are removed from the data. In addition, while scrubbing does not yield any further improvement, low-pass filtering at 0.20 Hz leads to a small improvement.

Keywords: aCompCor; physiological noise; fMRI artifacts; noise correction techniques; FIX; motion biases

1. Introduction

Functional connectivity (FC) using resting-state functional magnetic resonance imaging (fMRI) has attracted much attention since Bharat Biswal and colleagues first demonstrated that, during rest, the blood-oxygen-level-dependent (BOLD) signals in distinct areas of the somatomotor network are temporally correlated (Biswal et al., 1995). Strategies for studying resting-state FC have advanced in the last two decades allowing the identification of large-scale functional networks, termed resting-state networks (RSNs; Fox et al., 2005; Stephen M. Smith et al., 2013b). RSNs correspond to functional networks that activate on a range of tasks (Smith et al., 2009). While the spatial pattern of RSNs is similar across subjects, a recent study has demonstrated high accuracy in the identification of participants using FC estimates from repeated scans as fingerprints (Finn et al., 2015). Moreover, FC estimates have been shown to predict behavioral measures in individuals (Smith et al., 2015) while significant differences in FC have been reported in patients from a range of cerebrovascular and mental disorders compared to healthy subjects (Demirtaş et al., 2016; Leonardi et al., 2013; Woodward and Cascio, 2015). These findings suggest that FC has the potentials to improve our understanding regarding the functional organization over development, aging, and diseases states, as well as assist in the development of new biomarkers.

However, a main problem in fMRI is that significant variance on the BOLD signal is driven by head motion which has shown to cause severe consequences in FC studies (Power et al., 2015; Satterthwaite et al., 2019). Motion artifacts tend to be more similar in nearby regions compared to distant regions (Power et al., 2012; Satterthwaite et al., 2012; van Dijk et al., 2012). As a result, correlations between regions that are close to each other (short-distance correlations) tend to be inflated by motion more compared to distant regions (long-distance correlations) (see for example Fig. 5 in Satterthwaite et al., 2013). In addition, if a study compares differences in FC between populations that present different levels of motion and this is not accounted for in the preprocessing and analysis of the data, then the motion artifacts can cause artificial differences in FC between the examined populations. This phenomenon is particularly problematic for studies of development, aging and disease as children, elderly and patients tend to move more during the scan than young or control subjects (Power et al., 2015).

Importantly, confounds in fMRI arise also from physiological noise (Caballero-Gaudes and Reynolds, 2017; Murphy et al., 2013). Cardiac pulsatility in large vessels caused by cardiac-related pressure changes generates small movements in and around large vessels. In turn, these movements introduce fast pseudo-periodic fluctuations (~1 Hz) on the BOLD signal that are in phase with the cardiac cycle (Dagli et al., 1999). The high-frequency cardiac artifacts affect among others areas around the brainstem as well as areas in the superior sagittal sinus and lateral sulcus. On the other hand, breathing motion introduces high-frequency artifacts (~0.3 Hz) mostly at the edges of the brain. However, these are not the only sources of artifacts related to cardiac and breathing activity. Slow-frequency fluctuations in heart rate and breathing pattern (<0.1 Hz) are typically observed during rest and have a direct effect on the cerebral blood flow and levels of oxygenated hemoglobin in the brain (Birn et al., 2006; Chang et al., 2009; Kassinopoulos and Mitsis, 2019; Shmueli et al., 2007). As such they affect widespread regions in the gray matter (GM). Group-level statistical maps generated in our previous work with areas affected by the aforementioned physiological processes are available on https://neurovault.org/collections/5654/ (Fig. 12 in Kassinopoulos and Mitsis, 2019). Finally, widespread regions in GM are also prone to artifacts induced by slow spontaneous fluctuations in levels of arterial carbon dioxide (Prokopiou et al., 2019; Wise et al., 2004) and blood pressure (Whittaker et al., 2019). Therefore, physiological processes can considerably affect FC estimates if not taken into account during the preprocessing.

Several noise correction techniques (NCTs) have been proposed to correct for head motion and phyiological noise that can be classified as model-based or model-free techniques. In the case of head motion, model-based techniques are based on the motion parameters (MPs) estimated from volume realignment done at the very first steps of preprocessing. Three translational and three rotational displacement parameters are estimated from volume realignment that describe the rigid-body movement of head in space yielding in total 6 MPs. The most common practise used in FC studies to account for motion is to remove the 6 MPs from the data through linear regression (Power et al., 2015). Sometimes the derivatives of the 6 MPs or even the squared terms of these 12 time series are also removed from the data (Satterthwaite et al., 2013). Another practice employed in recent studies, named scrubbing, is to identify volumes contaminated by strong motion

artifacts and discard them from the data or replace them with values from the adjacent volumes using interpolation (Power et al., 2015).

With regards to physiological noise, model-based techniques utilize physiological recordings collected during the fMRI scan. Typically, the cardiac and breathing activity are recorded through a pulse oximeter and a respiratory bellow, and are used to model artifacts related to cardiac pulsatility and breathing motion with a technique named RETROICOR (Glover et al., 2000). RETROICOR uses the physiological recordings to generates nuisance regressors of cosines and sines that are in phase with the cardiac and breathing cycle. Subsequently, the extracted nuisance regressors are removed from the data through linear regression. In addition, the cardiac and breathing signals are often used to model fluctuations induced by changes in heart rate and breathing pattern. The heart rate and a respiratory measure such as the respiration volume per time are extracted from the physiological recordings and, subsequently, convolved with the so-called cardiac and respiration response functions. The outputs of these convolutions are used as nuisance regressors to account for the effect of heart rate and breathing pattern (Birn et al., 2008, 2006; Chang et al., 2009; Kassinopoulos and Mitsis, 2019).

An alternative option for noise correction in fMRI are model-free techniques that, in contrast to model-based techniques, do not require external physiological recordings and, have the theoretical benefit to be independent of a pre-established model. Some model-free techniques make use of principal component analysis (PCA) or independent component analysis (ICA) to decompose the fMRI data into a number of components (Behzadi et al., 2007; Perlbarg et al., 2007; Pruim et al., 2015; Salimi-Khorshidi et al., 2014). Then, components associated to noise are identified based on criteria such the spatial pattern or frequency content of a component, and their corresponding time series are subsequently removed from the data. Further, low-pass filtering at about 0.08 Hz is commonly used to remove high-frequency noise as RSNs are known to exhibit slow oscillations below 0.1 Hz (Damoiseaux et al., 2006). Finally, the mean time series across voxels in the whole brain, referred to as global signal (GS), as well as mean time series from voxels in the three tissue compartments, GM, white matter (WM) and cerebrospinal fluid (CSF), are sometimes considered as nuisance regressors to account for global artifacts (Power et al., 2017).

Recent studies attempted to compare the performance of a variety of NCTs as well as preprocessing pipelines that consist of a combination of techniques mentioned earlier (Birn et al., 2014; Burgess et al., 2016; Ciric et al., 2017; Parkes et al., 2018). A number of quality control (QC) metrics were used in these studies that reflect either the identifiability of RSNs or the mitigation of motion effects. However, a commom finding from many studies is that the scores obtained from the QC metrics for the examined NCTs or pipelines often yielded contradictory results. For example, while a pipeline would have been found to exhibit the highest score in terms of RSN identifiability it would have failed to reduce motion artifacts as good as other pipelines. Moreover, even though there is strong evidence that model-free techniques based on PCA or ICA are able to reduce artifacts due to head motion or physiological noise (Behzadi et al., 2007; Muschelli et al., 2014; Pruim et al., 2015; Salimi-Khorshidi et al., 2014), it is still an open question whether combining them with model-based techniques can provide superior performance.

In this work, we examined the performance of model-free and model-based techniques using QC metrics previously proposed and novel metrics related to large-scale network identifiability and presence of motion artifacts and biases. Multisession resting-state fMRI data from the Human Connectome Project were considered (Van Essen et al., 2013). With respect to model-free approaches, we examined FIX ("FMRIB's ICA-based X-noisefier"; Salimi-Khorshidi et al., 2014) as well as variants of aCompCor. FIX consists of whole-brain ICA decomposition followed by removal of noisy components identified using a multi-level classifier (Salimi-Khorshidi et al., 2014). Anatomical CompCor (aCompCor) refers to removal of the first five principal components from two noise regions of interest (ROIs), namely the WM and CSF compartments (Behzadi et al., 2007). Here, apart from evaluating the performance of the original aCompCor approach, we sought to answer whether removing more components would be beneficial examining components from WM and CSF separately. Finally, for the variant of aCompCor that exhibited the best improvement in QC scores, we investigated the additional benefit of removing nuisance regressors derived from the MPs and physiological recordings, excluding motion-contaminated volumes from the analysis and doing low-pass filtering before the removal of regressors.

2. Methodology

Unless stated otherwise, the preprocessing and analysis described below were performed in Matlab (R2018b; Mathworks, Natick MA).

2.1 Human Connectome Project (HCP) Dataset

We used resting-state scans from the HCP S1200 release (Glasser et al., 2016; Van Essen et al., 2013). The HCP dataset includes, among others, resting-state (eyes-open and fixation on a cross-hair) data from healthy young (age range: 22-35 years) individuals acquired on two different days. On each day, two 15-minute scans were collected. We refer to the two scans collected on days 1 and 2 as R1a/R1b and R2a/R2b, respectively. fMRI acquisition was performed with a multiband factor of 8, spatial resolution of 2 mm isotropic voxels, and a repetition time TR of 0.72 s (Glasser et al., 2013).

The minimal preprocessing pipeline for the resting-state HCP dataset is described in Glasser et al. (2013). In brief, the pipeline includes gradient-nonlinearity-induced distortion correction, motion correction, EPI image distortion correction, non-linear registration to MNI space and mild high-pass (2000 s) temporal filtering. The motion parameters are included in the database for further correction of motion artifacts. Apart from the minimally preprocessed data, the HCP provides a cleaned version of the data whereby time series corresponding to ICA components that FIX classified as noisy as well as 24 motion-related regressors (i.e., the 6 MPs estimated during volume realignment along with their temporal derivatives and the squared terms of these 12 time series) were regressed out of the data (Smith et al., 2013). The cleaned fMRI data are referred to later as FIX-denoised data.

Both minimally-preprocessed and FIX-denoised data are available in volumetric MNI152 and grayordinate space. The grayordinate space combines cortical surface time series and subcortical volume time series from GM, and has more accurate spatial correspondence across subjects than volumetrically aligned data (Glasser et al., 2013), particularly when the fMRI data have high spatial resolution such as in HCP.

In the present work, we examined minimally-preprocessed and FIX-denoised data in both volumetric and grayordinate space, from 390 subjects which included good quality physiological signals (cardiac and breathing waveforms) in all four scans, as assessed by visual inspection. The cardiac and breathing signals were collected with a photoplethysmograph and breathing bellow respectively.

2.2 Parcellation of the fMRI data

The following three fMRI-based atlases were considered in this study:

a. The Gordon atlas (Gordon et al., 2016). This atlas is composed of 333 cortical regions with 285 parcels belonging to one of twelve large-scale networks while the remaining ones are unassigned. Only the 285 parcels that are assigned to networks were considered in this study.
b. The Seitzman atlas (Seitzman et al., 2018) This atlas consists of 239 cortical, 34 subcortical and 27 cerebellar volumetric parcels. Among the 300 parcels, 285 parcels are assigned to one of thirteen large-scale networks and only these ones were considered here.
c. The MIST atlas: The MIST atlas is available in several resolutions ranging from 7 parcels to 444 parcels. In this study, we considered the MIST_444 parcellation that consists of 444 regions from the whole brain that are assigned to the 7 networks of MIST_7 parcellation.

All three atlases were defined on resting-state fMRI data and all (MIST) or the majority of (Gordon and Seitzman) their parcels are assigned to large-scale networks. The association of the parcels to networks is required for three of the quality control (QC) metrics described later (i.e., FCC, FD-FCC and ICCC). Therefore, as mentioned earlier, parcels that do not belong to a specific network were excluded from the study.

Before the parcellation, in the case of the Seitzman atlas, we performed spatial smoothing on the fMRI data with a Gaussian smoothing kernel of 5 mm full width half maximum (FWHM). Spatial smoothing is commonly done on fMRI data to suppress spatially random noise and enhance the signal to noise ratio (SNR). However, when mapping the fMRI data to a parcellation with relatively large parcels such as the parcels in the Gordon and MIST atlases, spatial smoothing is implicitly done. Therefore, we chose to omit spatial smoothing for these two atlases. We did spatial smoothing before conducting the mapping to the Seitzman atlas because this atlas consists of small spherical ROIs of 8 mm diameter (Seitzman et al., 2018) and, thus, if spatial smoothing is not performed, the parcel time series extracted from these ROIs may suffer from low SNR.

To speed up the preprocessing step, the regression of the nuisance regressors for each pipeline was performed in a parcel- rather than voxel-wise manner. In other words, the minimally-preprocessed and FIX-denoised fMRI data were first mapped to parcel time series by averaging the time series of all voxels or surfaces falling within a parcel and, subsequently, the resulted parcel time series were corrected for noise using the techniques described later. While it is significantly faster, this procedure yields same results with the steps done in the reversed order (i.e., first, the data are preprocessed and, then, are mapped to the associated parcels). The mapping of the fMRI data to the Gordon parcel space was done using the fMRI data in the grayordinate form while the mapping to the Seitzman and MIST parcel space was done using the volumetric form of the fMRI data. Finally, all parcel time series were high-pass filtered at 0.008 Hz.

2.3 Tissue-based regressors

The T1-weighted (T1w) images of each subject are provided in the HCP database in both native and MNI152 space. To extract the tissue-based regressors used by several pipelines examined here, first we performed tissue segmentation on the T1w images at the MNI152 space using FLIRT in FSL 5.0.9 that generated probabilistic maps for the GM, WM and CSF compartments (Zhang et al., 2001). Subsequently, the GM, WM and CSF binary masks were defined as follows: if a voxel had a probability above 0.25 to belong to GM then it was assigned as voxel in GM, if the probability for WM was above 0.9 it was assigned as voxel in WM, while if the probability for CSF was above 0.8 it was assigned as voxel in CSF. The choice of the threshold values was done based on visual inspection while overlaying the binary maps on the T1w images. Finally, based on these maps, the global signal was defined as the mean fMRI time series across all voxels within GM. In addition, PCA regressors were obtained separately for voxels in WM and CSF. The GS and PCA regressors were derived from both the minimally-preprocessed and FIX-denoised fMRI data in the volumetric space.

2.4 Model-based regressors related to motion and physiological fluctuations

A main goal of this study was to quantify the effect of model-based NCTs with respect to the quality of the fMRI data for atlas-based FC analysis and how they contribute to fMRI denoising when combined with tissue-based regressors. Therefore, for each scan the following four sets of model-based regressors were considered:

a. **Motion parameters (MPs):** The 6 MPs derived from the volume realignment during the minimally-preprocessing are provided in the HCP database as well as their temporal derivatives. In addition to the 12 aforementioned time series (12 MPs), we derived their squared terms, yielding in total 24 motion parameters (24 MPs).
b. **Cardiac regressors:** The cardiac regressors were modelled with 3rd order RETROICOR (Glover et al., 2000) using the cardiac signal of each scan. The regressors aimed at accounting for the effect of cardiac pulsatility on the fMRI time series.
c. **Breathing regressors:** The breathing regressors were modelled with 3rd order RETROICOR using the breathing signal of each scan (Glover et al., 2000). The regressors aimed at accounting for the effect of breathing motion.
d. **Systemic low frequency oscillations (SLFOs):** The SLFOs refer to non-neuronal physiological BOLD fluctuations induced by changes in heart rate and breathing pattern which were modelled following the

framework proposed in our previous work (Kassinopoulos and Mitsis, 2019). The heart rate and respiratory flow extracted from the cardiac and breathing signals for each scan were convolved with scan-specific cardiac and respiratory response functions and the outputs of these convolutions were linearly combined to model the SLFOs. To estimate the scan-specific physiological response functions and determine the linear combination of heart rate- and respiratory flow- related components needed to model the SLFOs, numerical optimization techniques were employed that maximize the fit of the model output (i.e., the SLFOs-related time series) to the GS. The GS was used as a fitting target as it is strongly driven by fluctuations in heart rate and breathing pattern. For more information on this method we refer the reader to Kassinopoulos and Mitsis (2019). The codes for the estimation of SLFOs can be found on https://github.com/mkassinopoulos/PRF_estimation.

Even though we selected subjects with good quality of physiological recordings, it was still important to preprocess both the cardiac and breathing signal to ensure the extraction of good heart rate and respiratory flow traces. To this end, we applied temporal filtering and correction for outliers as described in (Kassinopoulos and Mitsis, 2019). Moreover, as the effect of HR and breathing pattern variations on the fMRI BOLD signal is considered to last about half a minute (Kassinopoulos and Mitsis, 2019) physiological recordings for at least half a minute before the beginning of the fMRI acquisition would be required to account for these effects. However, due to the lack of data at this period, the first 40 image volumes were disregarded from the fMRI data, while the corresponding physiological data were retained so that the SLFOs could be modelled from the beginning of the considered fMRI scan.

2.5 Framewise data quality indices

A common index of quality in fMRI data is the framewise displacement (FD) introduced by Power et al. (2012). This index is defined as the sum of absolute values of the first derivatives of the 6 motion (realignment) parameters, after converting the rotational parameters to translational displacements on a sphere of radius 50 mm. FD is essentially a time series that reflects the extent of motion during the scan. In this work, FD was used for six QC metrics that are described in Section 2.7 to quantify the degree of motion artifacts and biases in preprocessed data. In addition, it was used to examine the effect of scrubbing which is the process whereby volumes associated with relatively large FD values are discarded before any further analysis (Section 2.6.4).

Another widely used framewise index of data quality is DVARS (Derivative of rms VARiance over voxelS) which measures how much the intensity of an fMRI volume varies at each timepoint compared to the previous point (Power et al., 2012). DVARS is defined as the spatial root mean square of the voxel time series after the time series are temporally differentiated. While DVARS is obtained from the fMRI time series and is not directly related to head movement, demonstrates similar traces with FD (Power et al., 2012). In similar to FD, DVARS was used in this study for two QC metrics related to the effect of motion as well as to flag volumes corrupted by motion.

2.6 Noise correction techniques (NCTs)

In this work, we assessed the performance of a large number of preprocessing pipelines using nine QC metrics that quantify the improvement in network identifiability and reduction of motion-related artifacts and biases. The pipelines consisted of commonly used preprocessing strategies, namely scrubbing, temporal low-pass filtering and removal of nuisance regressors through linear regression. To better understand the effect of each of the aforementioned strategies, five different groups of pipelines were examined that are described in the following sections. The QC metrics used to evaluate the performance of each pipeline are described in Section 2.7.

2.6.1 Optimizing aCompCor

In aCompCor, PCA regressors are obtained from the WM and CSF tissues and the first five components ordered by the variance explained in the WM and CSF voxel time series are used as nuisance regressors. This practice implicitly suggests that the PCA regressors that explain most of the variance in WM and CSF are also the ones with stronger association to model-based nuisance regressors. To examine whether the latter is the case, we estimated the variance explained on each PCA regressor from WM and CSF with the 24 MPs, the breathing regressors, the cardiac regressors, the SLFOs as well as all the aforementioned regressors combined. In addition, we examined the variance explained on each of the model-based regressors from a varying number of WM or CSF regressors of 1 to 100. The estimated explained variances corresponding to each model-based regressor were averaged across regressors of the same source of noise.

After confirming the hypothesis stated above, a main objective of this study was to examine the performance of WM and CSF denoising independently and determine the number of regressors that should be considered in the preprocessing to improve the quality of the fMRI data. To this end, for both noise ROIs, we considered the removal of the most significant PCA regressors with or without including the GS as an additional nuisance regressor with varying number of PCA regressors from 1 to 600 in a base 10 logarithmic base. Note that each scan consisted of 1160 fMRI volumes, therefore 600 components would correspond to about half of the available PCA regressors. Regarding the tissue-based regressors related to aCompCor, we refer to a set of regressors from WM and CSF as $WM^{x}_{(GS)}$ and $CSF^{x}_{(GS)}$, respectively, where x indicates the number of PCA regressors considered from each of the two tissue compartments and the presence of the string GS as superscript denotes the inclusion of the GS in the set of nuisance regressors. For example, the set of regressors WM^{200}_{GS} refers to the set consisting of the GS and the first 200 PCA regressors from WM. Note that the set WM^{200}_{GS} demonstrated the highest improvement in QC scores and, thus, the subsequent analyses in this work investigate the possibility of further improvement using additional strategies in the preprocessing along with the regression of this set.

2.6.2 Evaluation of data-driven NCTs employed in previous studies

Typically, fMRI studies consider only data-driven regressors for the preprocessing of the data that can be a combination of motion, tissue-based regressors or whole-brain component regressors (e.g., FIX). However, the number and kind of regressors included in the preprocessing vary across studies which raises the question whether all the pipelines perform equally well or some pipelines are more efficient compared to other ones. A selection of pipelines used in the literature were evaluated here using the QC metrics described in Section 2.7 to allow a comparison between them (Table 1). However, as the focus of this analysis was to examine the effect of the regressors per se rather than the entire pipeline, steps such as scrubbing (i.e., removal of motion-contaminated volumes) or temporal low-pass filtering were omitted. In addition, several pipelines were considered in this analysis that consisted of a small number of regressors (e.g. pipelines 1-5), even though, typically, more aggressive pipelines are found in the literature. These pipelines were considered in order to better understand possible differences in QC scores between more aggressive pipelines (e.g. pipelines 7 and 8).

Regarding the pipelines that involve FIX (i.e., pipelines 11-13), even though HCP database provides the results from MELODIC-ICA and, thus, we could remove noisy ICA components and further nuisance regressors from the minimally-preprocessed fMRI data in one step, we chose to remove the additional nuisance regressors from the FIX-denoised data found in the HCP database to be consistent with the approach taken in previous studies (Burgess et al., 2016; Siegel et al., 2017). Note that, for pipelines 12 and 13, we used the GS and WM/CSF regressors derived from the FIX-denoised data. Furthermore, as mentioned earlier, the FIX denoising performed in HCP included the removal of the 24 MPs.

Table 1. *Preprocessing pipelines based on data-driven approaches*

Pipeline	Sets of nuisance regressors considered in the pipeline	(Related to) pipeline used in:
1	6 motion parameters (6 MPs)	—
2	12 motion parameters (12 MPs; i.e., 6 MPs plus their derivatives)	—
3	24 motion parameters (24 MPs; i.e., 12 MPs plus their squared terms)	—
4	Global signal (GS)	—
5	12 MPs, GS	—
6	WM^5, CSF^5 (i.e., first 5 PCA regressors from white matter (WM) and first 5 PCA regressors from cerebrospinal fluid (CSF))	Behzadi et al., 2007
7	12 MPs, mean WM time-series (WM_{mean}) and mean CSF time-series (CSF_{mean})	Urchs et al., 2017
8	12 MPs, WM_{mean}, CSF_{mean}, GS	Finn et al., 2015
9	24 MPs, [GS, WM_{mean}, CSF_{mean}, plus their derivatives]	Laumann et al., 2017
10	24 MPs, [GS, WM_{mean}, CSF_{mean}, plus their derivatives and the squared terms of the 6 aforementioned time-series]	Ciric et al., 2016; Xia et al., 2018
11	FIX (i.e., the FIX-denoised data as provided from HCP)	Bijsterbosch et al., 2017; Smith et al., 2015; Zhang et al., 2018
12	FIX_{GS} (i.e., FIX followed by GS regression)	Burgess et al., 2016
13	FIX, GS, WM^5, CSF^5	Siegel et al., 2017
14	PCA regressors needed to explain 50% of variance in WM and CSF	Muschelli et al., 2014
15	GS, WM regressors needed to explain 30% of variance	—
16	GS, WM regressors needed to explain 35% of variance	—
17	GS, WM regressors needed to explain 40% of variance	—
18	GS, WM regressors needed to explain 45% of variance	—
19	GS, WM regressors needed to explain 50% of variance	—
20	WM_{GS}^{200} (i.e., the GS and the first 200 PCA regressors from WM)	—

2.6.3 Evaluation of model-based (motion and physiological) NCTs

Even though the motion parameters are indirectly derived from the data through the process of volume realignment, they do not purely correspond to motion-induced BOLD fluctuations rather than to rigid-body displacements. Therefore, treating them as nuisance regressors during the preprocessing inherently imposes some assumptions about the effect of motion on the BOLD signal which may not be valid. Similarly, the efficiency of physiological regressors that are obtained from concurrent physiological recordings (e.g. cardiac and breathing signals) depends on the validity of the models used to be estimated as well as on the quality of the recordings. Thus, an important question that needs to be addressed is whether the aforementioned model-based regressors contribute to the denoising of the fMRI data, and particularly when combined with tissue-based regressors that do not have the limitations of the model-based approaches. To this end, using the QC metrics, we evaluated 64 combinations of pipelines that employ sets of model-based and tissue-based regressors. Specifically, we considered as model-based regressors the 24 MPs, the cardiac and breathing regressors and the SLFOs regressor, while as tissue-based regressors we considered the GS and 200 PCA regressors from WM (WM_{200}).

2.6.4 Scrubbing

In the analyses preceding scrubbing, it was found that the set of nuisance regressors WM_{GS}^{200} yielded the highest QC scores. Therefore, the next question that we aimed to address is whether scrubbing can provide any further improvement in the QC scores for this specific set of regressors and at what scrubbing threshold. This analysis was done both using the FD and the DVARS to determine the motion-contaminated volumes that would be discarded. In the case with the FD, we repeated the analysis for the values of threshold FD_{thr} 0.15, 0.20, 0.25, 0.30, 0.50, 0.80 and 1.00 mm, whereas in the case with the DVARS the values of threshold $DVARS_{thr}$ 0.5, 1, 1.5, 2, 5, 10 and 20 median absolute deviations (MAD) were considered.

2.6.5 Low-pass filtering (LPF)

Similar to the analysis with scrubbing, we investigated whether low-pass filtering the data and the set of nuisance regressors WM_{GS}^{200} before their removal would yield higher QC scores compared to not doing low-pass filtering. The QC scores were estimated for the following values of cut-off frequency: 0.05, 0.08, 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 Hz.

2.7 Quality control (QC) metrics

Nine QC metrics that are described below were used to evaluate the data quality with respect to the identifiability of large-scale networks and presence of motion-related artifacts and biases. Pearson correlation was calculated between the time series of each pair of parcels resulting in an FC matrix per scan, pipeline and atlas. Apart from FDDVARS and FD-FDDVARS, all other metrics are based on the FC matrix. Note that throughout the text we refer to a pair of parcels as *edge*.

Functional connectivity contrast (FCC):

A main property of the three parcellations used in this study is that each parcel is assigned to a specific large-scale network which implicitly suggests that on average a pair of parcels belonging to the same network, also called within-network edge (WNE), should exhibit a higher correlation value compared to a pair of parcels from different networks (between-networks edge; BNE). Based on this property, we assumed that if a pipeline improves the signal-to-noise ratio (SNR) in the data this would lead to an even larger difference between correlation values of WNEs and BNEs. To quantify the extent to which WNEs had higher correlation values than BNEs after applying a preprocessing pipeline on the data, we used the metric FC contrast (FCC) defined as the z-statistic of the Wilcoxon rank-sum test related to the null hypothesis that WNEs and BNEs in the FC matrix are samples from continuous distributions with equal medians. In other words, the higher was the value of FCC the higher were the correlation values of WNEs compared to values of the BNEs. Furthermore, for the optimal pipeline found in this work, we quantified the identifiability of each of the networks separately using the same metric but considering only WNEs belonging to the network of interest rather than WNEs from all networks when comparing WNEs with BNEs.

FD-FCC:

While it is desired to improve the FCC score for the data of each scan, at the same time it is desired that low-motion scans and high-motion scans demonstrate similar FCC scores. Therefore, FD-FCC was defined as the correlation between the mean FD and FCC across scans and was used in this work to assess potential biases due to different levels of motion between scans.

Median of Intraclass correlation values (MICC):

ICC is a widely used metric in statistics to assess how reproducible measurements of the same quantity are across different observers or instruments (Shrout and Fleiss, 1979). Similar to previous studies evaluating the performance of preprocessing strategies in fMRI, we have used ICC to assess test-retest reliability across the four sessions of each subject in whole-brain FC estimates (Birn et al., 2014; Parkes et al., 2018; Shirer et al., 2015). For a pair of parcels i and j, $ICC_{i,j}$ was defined as

$$ICC_{i,j} = \frac{MS_b - MS_w}{MS_b + (k-1)\, MS_w}, \qquad [1.]$$

where k is the number of scans per subject (4), MS_b is the between-subject mean square of correlation values between parcels i and j, and MS_w is the within-subject mean square of correlation values for the same pair of parcels. The MICC score assigned to each pipeline for the assessment of data quality was defined as the median of $ICC_{i,j}$ across all edges.

ICC contrast (ICCC):

A common finding from previous studies is that MICC drops when a relatively aggressive pipeline is used (Birn et al., 2014; Parkes et al., 2018; Shirer et al., 2015). This finding suggests that artifacts in fMRI data have high subject-specificity and, thus, when the artifacts are reduced after the preprocessing, MICC decreases as well. As this metric was not very helpful in previous studies (Birn et al., 2014; Parkes et al., 2018), apart from MICC, we examined an extension of this metric named ICC contrast (ICCC) that measures how much higher are the $ICC_{i,j}$ values for the WNEs compared to the values for the BNEs. The assumption behind ICCC is that only WNEs should demonstrate high subject specificity. In similar to FCC, ICCC was defined as the z-statistic of the Wilcoxon rank-sum test related to the null hypothesis that

WNEs and BNEs in the ICC matrix are samples from continuous distributions with equal medians. In other words, the higher was the value of ICCC the higher were the $ICC_{i,j}$ values of WNEs compared to the values of BNEs.

FDDVARS:

To assess the presence of motion artifacts on the parcel time series after preprocessing, we calculated the Pearson's correlation between FD and DVARS (Muschelli et al., 2014). Note that while FD was estimated only once based on the motion (realignment) parameters, DVARS was estimated for each pipeline separately using the parcel time series after the preprocessing. The DVARS used for this metric was defined as the framewise data quality index DVARS described in Session 2.5 with the difference that it was estimated based on the parcel instead of voxel time series. The FDDVARS score assigned to each pipeline was obtained by averaging the estimated FDDVARS across all scans.

FD-FDDVARS:

While the FDDVARS score reflects the extent to which motion artifacts are present in the data, a low value of FDDVARS does not necessarily mean that the biases in the FC estimates due to motion are also low. High-motion scans are contaminated by more severe motion artifacts compared to low-motion scans which has been shown to systematically bias the estimated FC matrices (Power et al., 2015). Even though a preprocessing strategy may reduce the motion artifacts to both high- and low-motion scans, if in the preprocessed data there are still differences in the levels of motion artifacts between the two groups, this would result again in a systematic bias at the FC estimates of these groups. To assess the presence of motion-related biases, we used the QC metric FD-FDDVARS which was defined as the correlation between the mean FD and FDDVARS across scans.

$FDFC_{median}$:

For a pair of parcels i and j, $FDFC_{i,j}$ was defined as the correlation between the Pearson correlation of this pair (i.e., $FC_{i,j}$) and the mean FD across scans. To assess the quality of data with respect to motion-related biases in FC, each pipeline was assigned an $FDFC_{median}$ score that corresponded to the median absolute $FDFC_{i,j}$ across all edges (Parkes et al., 2018; Power et al., 2012).

$FDFC_{dist}$:

Early studies on the effect of motion in fMRI have shown that on raw data short-distance edges demonstrate stronger inflations in correlations due to motion than long-distance edges (Power et al., 2012; Satterthwaite et al., 2013). Based on these observations, to measure the degree of distance-dependent motion artifacts, we considered the QC metric $FDFC_{dist}$ which is defined as the correlation between the $FDFC_{i,j}$ value defined in the previous QC metric ($FDFC_{median}$) and the Euclidean distance that separates parcels i and j, across all edges (Ciric et al., 2017; Parkes et al., 2018).

FD – Mean FC (FD-MFC):

The metric $FDFC_{median}$ inherently assumes that the Pearson correlation of an edge is affected by motion in the same way across subjects. However, considering studies have reported differences in brain anatomy across subject (Bijsterbosch et al., 2018), we can assume that motion does not affect edge-wise FC estimates necessarily in the exact same way. Therefore, to assess the effect of motion on FC estimates in a looser manner, we propose the FD-MFC metric which is defined as follows: First, the mean of all Pearson correlations in the FC matrix (FCM) is estimated for each scan separately (considering only unique pairs of parcels) and, subsequently, the correlation between the mean FD and MFC across all scans is derived.

2.8 Normalization of QC metrics

The nine QC metrics described in Session 2.7 can be categorized into signal-related and motion-related metrics. The signal-related metrics being the FCC, MICC and ICCC, are meant to reflect the SNR of the data and are expected to yield low scores on data that have not been preprocessed as high levels of artifacts are likely to obscure the signal of interest. They are also expected to yield low scores on data that a very aggressive pipeline is applied and the signal of interest is lost. On the other hand, relatively high scores of signal-related metrics would be expected to be obtained from data whereby a good pipeline is applied and artifacts are reduced while the signal of interest is preserved. Motion-related metrics are expected to yield high absolute scores on data that have not been preprocessed indicating the presence of motion artifacts or biases whereas as we move to more aggressive pipelines, we expect these scores to approach zero reflecting the reduction of motion artifacts and biases.

As the goal of a preprocessing strategy is to remove artifacts while also preserving the signal of interest, ideally a pipeline that yields high scores in signal-related QC metrics and low scores in motion-related metrics would be preferred. However, due to that each QC metric is based on different assumptions and some metrics are based on Pearson correlation while other ones are based on the Wilcoxon rank-sum test, each metric illustrates a different trend across pipelines and yields a different range of scores (see for example Fig. 2) making the selection of the optimal pipeline difficult. Therefore, to overcome this drawback, we followed the following steps:

1. First, we randomly split the 390 subjects to 10 groups of 39 subjects ensuring that the groups were characterized by similar distributions of mean FD values.
2. Then, the QC scores were estimated for each of the 10 groups separately. Apart from MICC and ICCC, for all other metrics, only the first of the four scans were considered from each subject to avoid estimating correlations with repeated measures. MICC and ICCC were calculated using all four scans of each subject as by definition these two ICC-based metrics require repeated measures (scans) from each subject. FCC and FDDVARS that are calculated on a scan basis rather than within a group of scans were averaged across subjects within each group. That way, the quality of the data for a given atlas, pipeline and group of subjects was assigned with one score for each of the nice QC metrics.
3. Subsequently, all motion-related metrics were normalized to 1 – abs(x) where x is the score of each metric, so that, similarly to signal-related metrics, a high positive score is assigned to good quality data.
4. In the next step, the scores were expressed as z-scores based on the relation $z_{i,k,p} = \frac{x_{i,k,p} - \mu_i}{\sigma_i}$ where $x_{i,k,p}$ and $z_{i,k,p}$ are respectively the original and z-score values obtained for QC i, group of subjects k and pipeline p and, μ_i and σ_i are respectively the mean and standard deviations of the scores from all groups of subjects related to the QC i obtained from the raw data.
5. Subsequently, the z-scores of FCC and ICCC were averaged to yield the summarized score QC_{signal} and the z-scores of the 6 motion-related metrics FD-FCC, FDDVARS, FD-FDDVARS, $FDFC_{median}$, $FDFC_{dist}$ and FD-MFC were averaged to yield the summarized score QC_{motion}. The MICC was not included in the estimation of the QC_{signal} score as it was proved to reflect subject-specificity due to noise rather due to signal of interest as it was meant to.
6. Finally, the two latter summarized scores, QC_{signal} and QC_{motion}, were averaged to obtain the score for the combined quality control metric CQC.

The normalization described here allowed us to express each metric to z-scores that reflect relative improvement in standard deviations with respect to the raw data. Importantly, high z-scores can be interpreted as the associated QC metric exhibiting high sensitivity. That is, if the QC metric for a given pipeline exhibits high z-score it is very likely that in a new dataset the score for the same QC metric will be better when the data are preprocessed with the same pipeline compared to the raw data. Note that after the normalization of the QC metrics, all metrics were summarized into two indices, the QC_{signal} and the QC_{motion}, that, in turn, were averaged to obtain the final CQC score. However, while we present the results for both QC_{signal}, QC_{motion} and CQC, the choice for the best pipeline in each analysis is based on the third one.

3. Results

Here we present results mainly for the Gordon atlas since the results between the three examined atlases did no show significant differences. The results obtained using the Seitzman and MIST atlases can be found in the Supplementary Material.

3.1 Optimizing aCompCor

Fig. 1 shows the estimated variance explained with model-based regressors on each of the WM and CSF regressors (left panel) as well as the estimated variance explained from a set of WM and CSF regressors on the model-based regressors (right panel; for more information see Section 2.6.1). Note that the PCA regressors were ordered with decreasing variance explained on the tissue compartment they were originated from. As we see, the more variance a PCA regressor explained in the tissue compartment it originates from, a larger fraction of variance of that PCA regressor was explained by the model-based regressors. Moreover, the high-motion scans demonstrated different trends compared to low-motion scans. For example, looking at the left panel we see that the first PCA regressors demonstrated stronger association to the 24 MPs for the high-motion scans while stronger association to cardiac pulsatility was found for the low-motion scans.

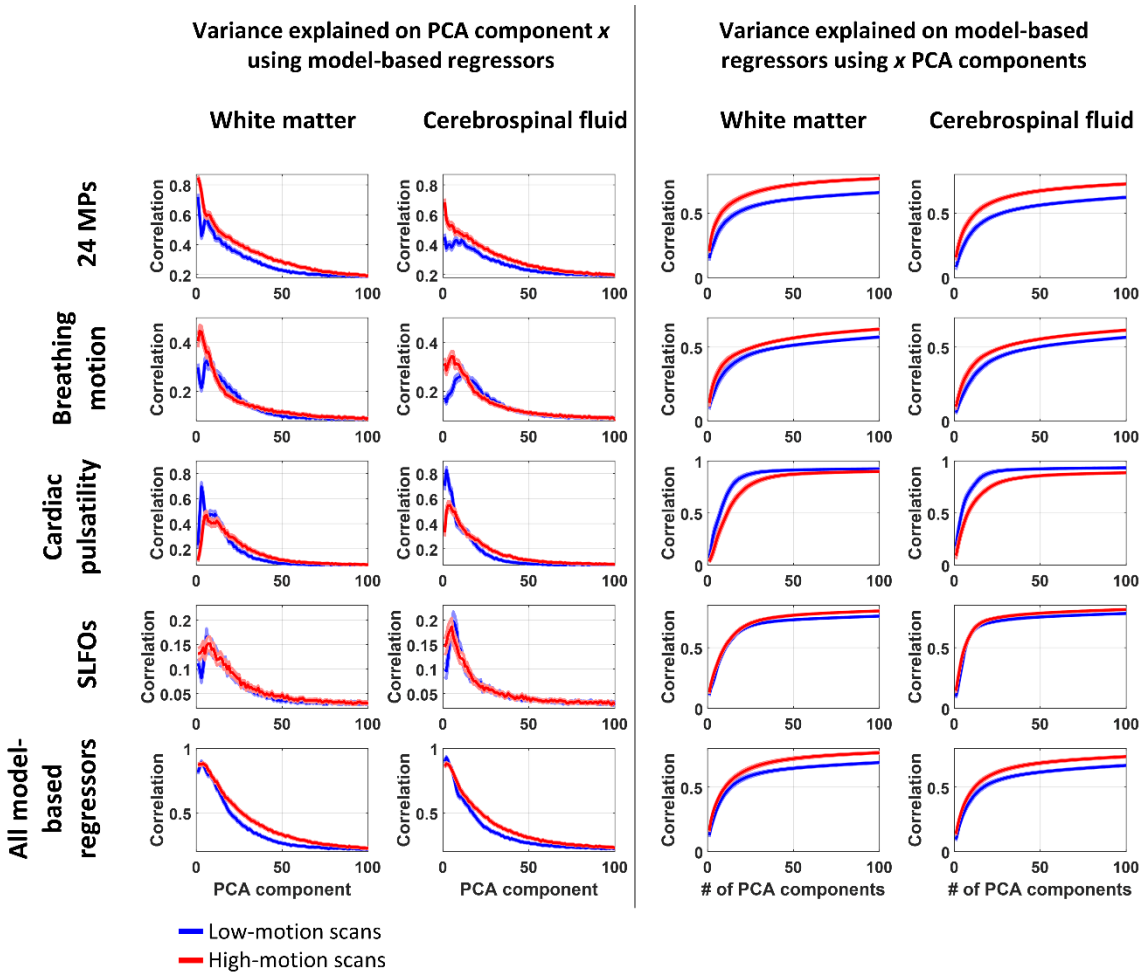


Fig. 1. Relation between model-based regressors and PCA regressors obtained from WM and CSF compartments. Left panel refers to the variance explained on each of the first 100 PCA components using the set of model-based regressors stated on the left of each row. Right panel refers to the mean variance explained on the regressors stated on the left of each row using a number of PCA components as explanatory variables indicated on the x axis. To examine the dependence of the curves on the degree of motion in each scan, two groups of scans were considered, referred to as low- and high-motion scans, that correspond respectively to the lower and upper quartile of the distribution of mean FD values. The blue and red curves correspond to the correlation averaged across low-motion and high-motion scans, respectively, whereas the shaded areas denote the standard error. For all four sources of noise, we observe that the first few PCA components demonstrated stronger association to the model-based regressors compared to components found later in the order justifying the practice of using the most significant PCA components in aCompCor.

Looking at WM vs CSF regressors, we observe several slight differences such as that the first few WM regressors explain better the 24 MPs compared to the CSF regressors, whereas the opposite is observed when looking at the cardiac pulsatility. However, when considering a large number of regressors (e.g. 100) both WM and CSF regressors explained significant fraction of variance in all four sets of model-based regressors with mean correlation values above 0.5 which suggests that both WM and CSF regressors can account to some extent for BOLD fluctuations due to head and breathing motion as well as cardiac pulsatility and SLFOs.

To determine the optimal number of PCA regressors that should be considered in the preprocessing we repeated the preprocessing with or without the GSR, including regressors either from WM or CSF and varying the number of PCA regressors. For each pipeline, we evaluated the quality of the preprocessed data using the QC metrics described in Session 2.7. Different trends were observed between the nine QC metrics varying the number of components considered (Fig. 2) which made the identification of an optimal pipeline difficult. To address this, we proceeded with the normalization of the metrics to z-scores as described in Session 2.8 (Suppl. Fig. 1) that allowed us to compare the sensitivity between the QC metrics and also give more weighting to metrics with high sensitivity when determining the optimal pipeline.

In similar to previous studies, the signal-related metric median intraclass correlation (MICC) yielded high scores in the raw data whereas when preprocessing was performed, the more WM or CSF regressors were considered the lower the MICC score was (Fig. 2c; Birn et al., 2014; Parkes et al., 2018). This trend is believed to be due to that noise in fMRI is characterized by high subject specificity and, hence, removing the noise in more aggressive pipelines leads to reduction in subject specificity as well (Birn et al., 2014; Parkes et al., 2018). As the MICC metric did not seem to reflect the preservation of signal in the data as it was meant to be used for, we excluded it from the rest of the analysis.

Based on the z-scores, we summarized the results from the two signal-related metrics, FCC and ICCC, and six motion-related metrics, FD-FCC, FDDVARS, FD-FDDVARS, $FDFC_{median}$, $FDFC_{dist}$ and FD-MFC, to the metrics QC_{signal} and QC_{motion}, respectively, as shown in Fig. 3. Fig. 3 also shows the scores for the combined QC metric CQC which was defined as the average score between QC_{signal} and QC_{motion}. Note that as QC_{signal} and QC_{motion} were defined, the former reflects the enhancement of SNR in data whereas the latter reflects the reduction in motion artifacts and biases. In Fig. 3, we observe that even though WM and CSF denoising exhibited similar performance in terms of mitigating motion effects, WM denoising achieved considerably higher SNR compared to CSF denoising. Furthermore, including the GS to the nuisance regressors significantly improved the scores for both QC_{signal} and QC_{motion}, particularly for low number of PCA components. Due to these observations, the discussion is focused on the performance of WM denoising, and if it is not explicitly stated it is assumed that the GS is also included in the set of regressors.

Overall, we observe that QC_{signal} was high for the sets of regressors WM_{GS}^{30} to WM_{GS}^{200} with a maximum score of 8.9 for WM_{GS}^{60}. In contrast, QC_{motion} illustrated a sharp peak for the more aggressive set of regressors WM_{GS}^{200} and, as a consequence, the optimal set of regressors according to CQC was the latter one (i.e., WM_{GS}^{200}). Note that the fMRI scans considered in this study consisted of 1160 volumes, therefore the 200 WM regressors used in the preprocessing correspond to ~ 17% of the available WM regressors. The analysis of optimizing aCompCor was repeated for the data at the Seitzman and MIST parcel space and yielded similar trends for varying number of PCA regressors. Similar to the data in the Gordon space, the set WM_{GS}^{200} was found to be the best choice for the data in the Seitzman atlas space, whereas the set WM_{GS}^{300} seemed to perform slightly better for the data in the MIST atlas space. In the following analyses, for both three atlases, we considered WM_{GS}^{200} when comparing the performance with other preprocessing strategies (e.g. including model-based regressors or performing scrubbing and low-pass filtering before the regression of nuisance regressors).

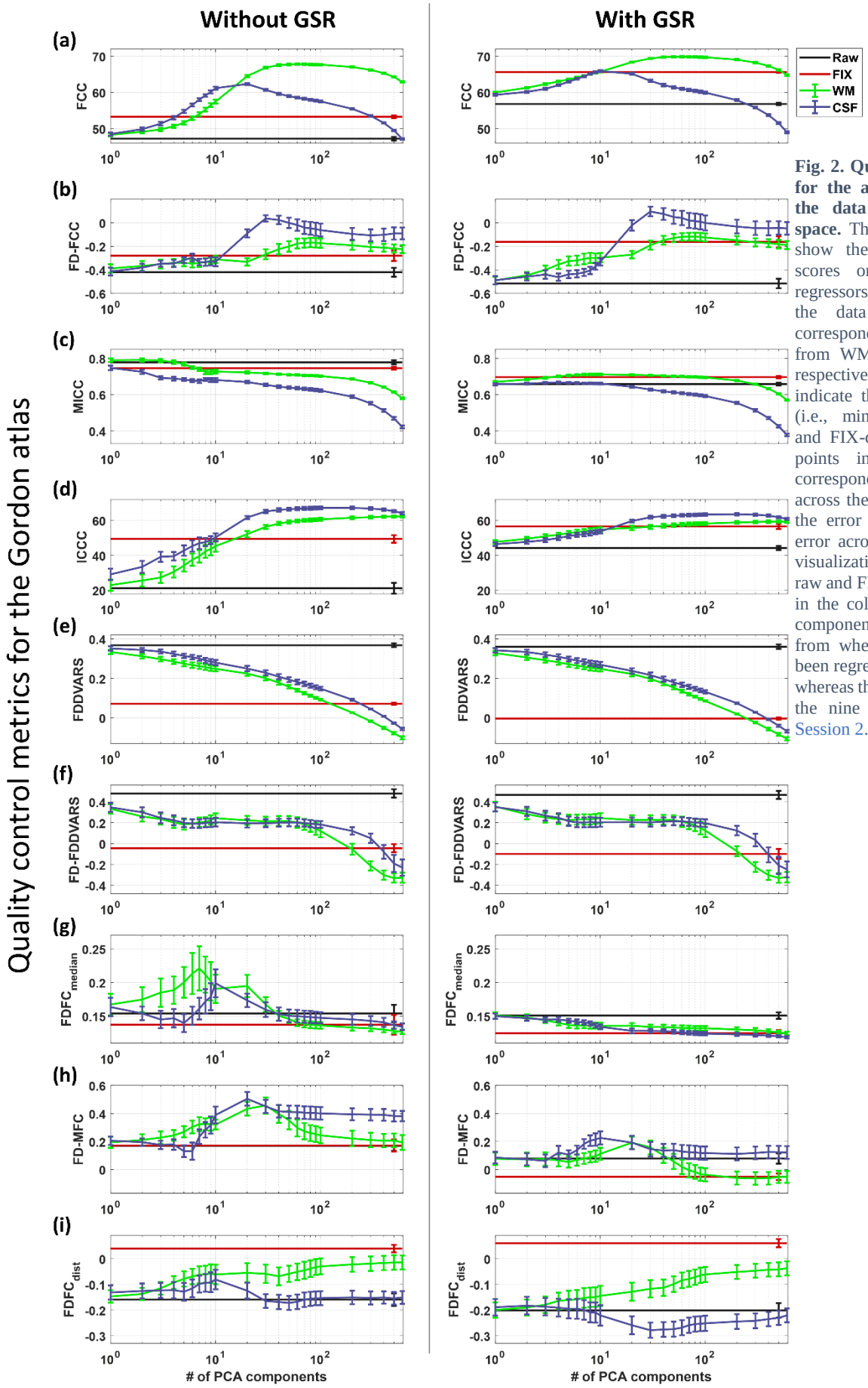


Fig. 2. Quality control (QC) scores for the aCompCor analysis using the data in the Gordon parcel space. The green and purple curves show the dependence of the QC scores on the number of PCA regressors that were removed from the data with the two colors corresponding to PCA regressors from WM and CSF compartments, respectively. The black and red lines indicate the QC scores for the raw (i.e., minimally-preprocessed data) and FIX-denoised data. The middle points in the curves and lines correspond to the QC scores averaged across the 10 groups of subjects and the error bars indicate the standard error across the 10 groups. To ease visualization, the error bars for the raw and FIX-denoised data are shown in the column corresponding to 500 components. The two columns differ from whether the global signal has been regressed out (right) or not (left) whereas the rows (a)-(i) correspond to the nine QC metrics described in Session 2.7.

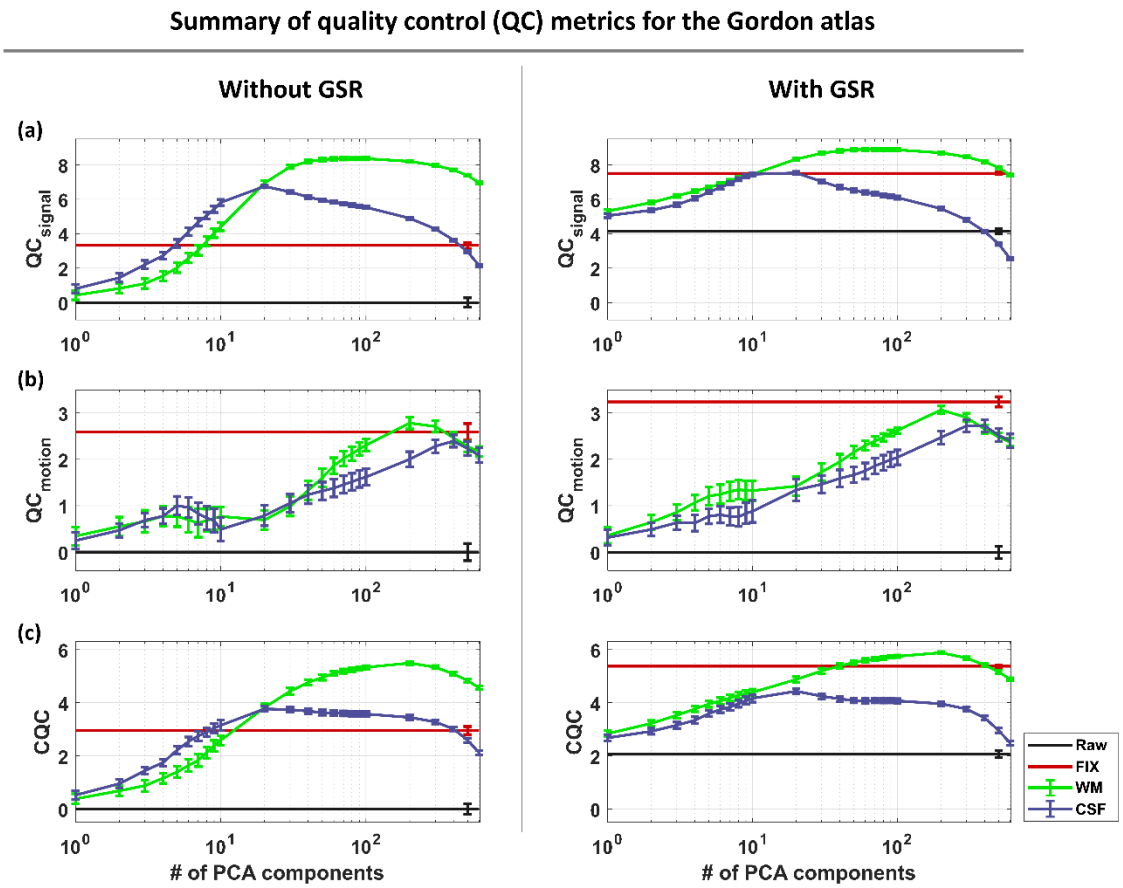


Fig. 3. Summarized quality control (QC) scores for the aCompCor analysis using the data in the Gordon parcel space. The z-scores of two signal-related metrics FCC and ICCC, and the six motor-related metrics FD-FCC, FDDVARS, FD-FDDVARS, $FDFC_{median}$, $FDFC_{dist}$ and FD-MFC were averaged to yield the summarized scores QC_{signal} (a) and QC_{motion} (b), respectively. Subsequently, the two latter summarized scores were averaged to obtain the combined QC metric (CQC). We observe that about 50 to 100 PCA regressors from WM were needed in order to achieve high score of QC_{signal} while 200 components from WM demonstrated the highest score in QC_{motion}. Including the GS in the set of regressors led to slightly higher scores for both summarized metrics. With respect to the CQC metric, the set of regressors WM_{GS}^{200} yielded the highest score (5.9) with the FIX_{GS} demonstrating the second highest score (5.4). While CSF denoising yielded as high scores as WM denoising with respect to reduction of motion artifacts and biases (QC_{motion}), it also led to loss of signal of interest based on the low scores of QC_{signal}.

Signal-related QC metrics

Functional connectivity contrast (FCC):

The metric FCC proposed in this work for assessing the identifiability of large-scale networks exhibited unimodal curves for both WM and CSF, both with and without GSR (Fig. 2a). However, WM denoising achieved higher scores with a maximum score of 69.9 for WM_{GS}^{60}. Fig. 4 shows the FC matrix for the raw (minimally-preprocessed) data and for data that have been preprocessed with different pipelines for a scan that demonstrated considerable improvement in identifiability of the networks when regressing out the set WM_{GS}^{200}. It is clear to see that the raw data were very noisy preventing the identification of the networks (FCC=23.0) but when denoising was conducted with WM_{GS}^{200}, all 12 networks were easily identified (FCC=69.4). Interestingly, when GSR was applied without any other nuisance regressor or NCT on the raw data, it did not have a strong effect on the contrast but when it was applied after FIX denoising it led to a significant increase of FCC score from 40.1 to 65.4. However, overall, GSR improved the FCC score for both FIX and WM denoising (Fig. 2a).

Fig. 5 shows the FC matrices averaged across all 1560 scans (group-level FC matrices) obtained from raw and four preprocessed fMRI datasets (i.e., data preprocessed with different pipelines). The FCC estimated from the group-level FC matrices were substantially higher compared to the FCC estimated on a scan basis for the same pipelines (Fig. 2a).

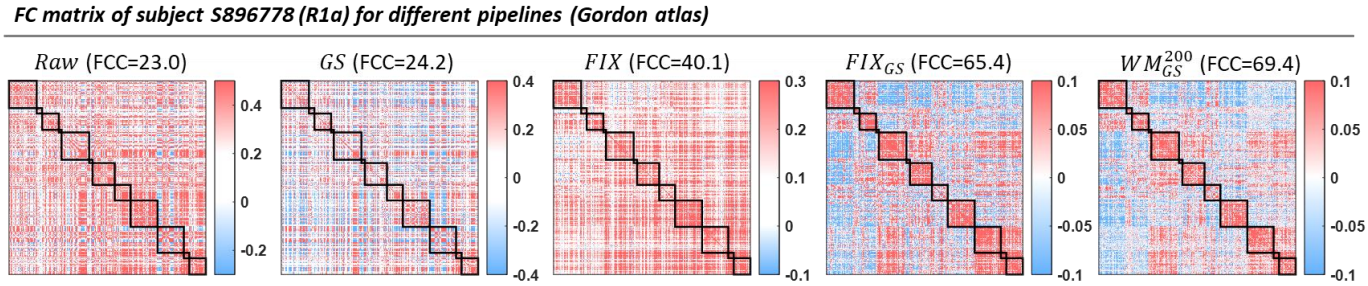


Fig. 4. FC matrix of subject S896778 (R1a) for different pipelines (Gordon atlas). While the networks could not be distinguished by visually inspecting the FC matrix of the raw data, they were easily identified after regressing out the set WM_{GS}^{200} or after FIX denoising, especially when FIX was combined with GSR. Similar conclusions were drawn based on the FCC metric that quantifies the identifiability of the networks (reported on the top of each matrix).

Note that for the raw data, the FCC score that was estimated first on a scan-basis and, then, averaged across all scans was 47.3 (Fig. 2a) whereas the FCC score estimated from the group-level FC matrix (i.e. the FC matrix was first averaged across all scans) had a higher value of 67.2 (Fig. 5). In addition, the FCC score obtained from the group-level FC matrix (67.2) was at similar levels with the highest FCC score achieved on a scan-basis across all pipelines (i.e., when preprocessed with WM_{GS}^{200}).

Intraclass correlation contrast (ICCC):

The metric ICCC proposed in this work to assess subject specificity in the fMRI data, showed an increasingly monotonic behavior for both WM and CSF (both with and without GSR), reaching a plateau at about 30 components with a small decline for the most aggressive pipelines. However, as we see in Suppl. Fig. 1 (a & d), FCC exhibited higher z-scores compared to ICCC and, therefore, contributed the most to the scores in the signal-related summarized metric QC_{signal} (Fig. 3a) which was defined as the average z-score between FCC and ICCC.

Fig. 5 shows the ICC matrices for the raw data and four preprocessed fMRI datasets estimated using all 1560 scans. As we can see, in raw data the ICC values were high for all edges which resulted to a low ICCC whereas when an aggressive pipeline was used (e.g. WM_{GS}^{200}) the ICC values for most of the BNEs dropped to significantly lower values compared to

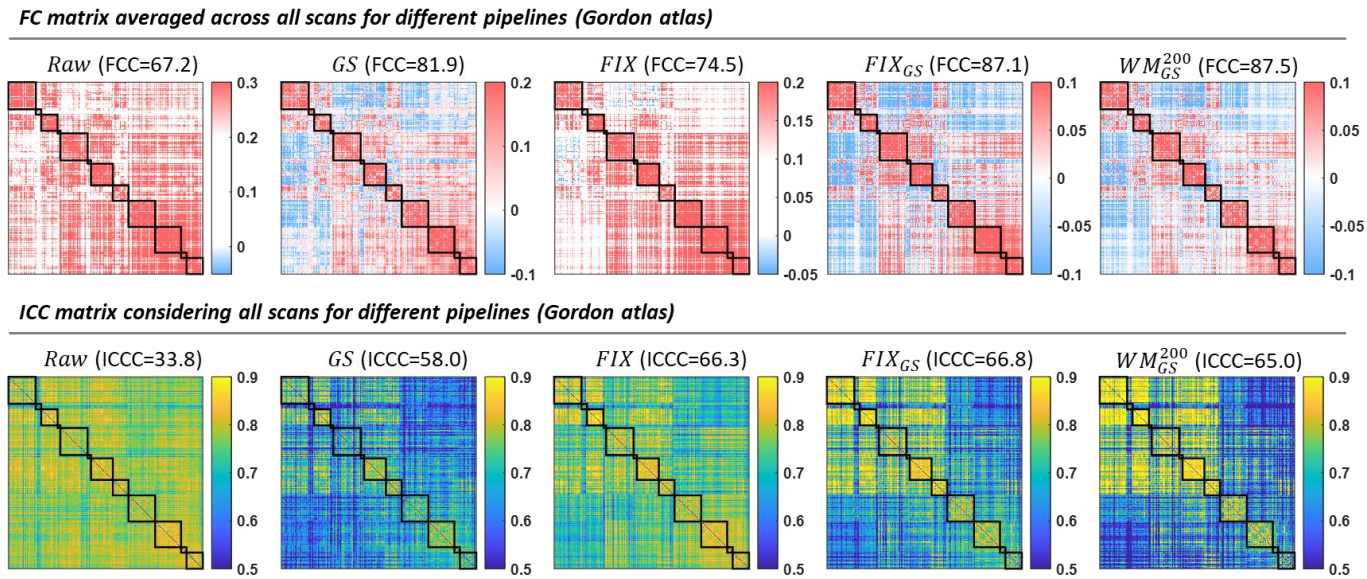


Fig. 5. FC (top) and ICC (bottom) matrices considering all scans for different pipelines obtained from the data in the Gordon parcel space. Averaging the FC matrices across all 1560 scans improved the identifiability of the networks considerably for both the raw and preprocessed data. As a consequence, the associated FCC scores reported on the top of each matrix are higher than the scores presented in Fig. 2a which were obtained on a scan-basis and then averaged within groups of 39 subjects. Similarly, the contrast estimated from the ICC matrices (i.e., ICCC) when considering all 1560 together was higher compared to the ICCC estimated from the smaller groups of 39 subjects each (Fig. 2d). Interestingly, we observe that a large number of BNEs, and especially edges between the default mode and fronto-pariental networks, exhibited low FC values but high ICC values.

the rest of the edges leading to an increase in ICCC score. Nevertheless, even with aggressive pipelines, many BNEs, and particularly edges corresponding to interactions between the default mode and fronto-parietal networks, demonstrated high ICC scores even though the corresponding edges in the group-level FC matrix showed low correlation values (Fig. 5). Similar results were observed for the Seitzman and MIST atlas as well (Suppl. Fig. 6-Suppl**. Fig. 7**). In addition, note that the ICCC scores reported in Fig. 5 are higher compared to the ICCC scores extracted from the smaller groups of subjects shown in Fig. 2d (groups of 39 subjects each). Also, differences in ICCC between pipelines found when scores were obtained for each group of subjects separately were decreased when ICCC was obtained from all subjects in one step (e.g. differences in ICCC scores between GS and FIX_{GS}). The aforementioned property of the metric ICCC suggests that the sensitivity of this metric in comparing the performance between preprocessing strategies decreases when larger number of subjects is considered.

Motion-related QC metrics

FD-FCC

As can be seen in Fig. 2b, the raw data yielded a mean FD-FCC score of -0.42 implying that the lower were the levels of motion in a scan the higher the FCC score was. Importantly, when performing WM denoising with more than 30 components, the strength of FC-FCC dropped to about -0.15 (z-score 2.2). Fig. 6 shows scatterplots of mean FD vs FCC for the first scan of 370 subjects (since a sufficient number of scans was available for this analysis, 20 subjects that demonstrated a mean FD three median absolute deviations (MADs) above the median were excluded). Note that the correlations of the scatterplots correspond to the FD-FCC scores for each pipeline and they only differ from the scores shown in Fig. 2b in that they were estimated in a single step using the first scan from all subjects whereas in Fig. 2b the correlation was estimated for each of the 10 groups of subjects separately considering again only the first scan of each subject. In Fig. 6 we observe that even though GSR applied alone on the raw data improved the FCC score, at the same time it increased the negative correlation between mean FD and FCC or, in other words, it enhanced the dependence of FCC score on the levels of motion ($r = -0.46$; $p = 10^{-19}$). However, when GSR was done along with WM denoising of 200 regressors (WM_{GS}^{200}) the negative correlation of FD-FCC was almost vanished ($r = -0.11$; $p < 0.04$).

FDDVARS

In Fig. 2e we see that the raw data demonstrated an FDDVARS score of 0.37 suggesting that the parcel time series were strongly contaminated by motion artifacts. WM denoising with 200 components (WM_{GS}^{200}) was able to drop FDDVARS to 0.02 which corresponded to a z-score of 11.3. Note that FDDVARS exhibited significantly higher z-scores than the rest of the motion-related QC metrics (Suppl. Fig. 1e) and, therefore, contributed the most to the scores of the summarized metric QC_{motion} (Fig. 3).

FD-FDDVARS

The raw data exhibited a mean FD-FDDVARS of 0.48 (Fig. 2f) implying that the higher were the levels of motion in a scan the stronger were the motion artifacts in the fMRI data. The set of regressors WM_{GS}^{200} achieved the smallest absolute score of FD-FDDVARS (score: -0.07) which corresponded to a z-score of 2.7. Fig. 6 shows scatterplots of the mean FD vs the FDDVARS score (i.e., FD-FDDVARS) for the raw data and four different preprocessed datasets considering the first scan from 370 subjects (20 subjects were excluded due to extreme values in mean FD). As we can see from the raw data, based on mean FD, the levels of motion during a scan had a strong effect on FDDVARS which reflects the degree of motion artifacts in the fMRI data ($r = 0.47$; $p < 10^{-21}$). The pipelines FIX_{GS} and WM_{GS}^{200} were able to reduce the score of FD-FDVARS to -0.18 ($p < 0.001$). We also observe that FIX achieved a lower FD-FDDVARS score of -0.10 compared to FIX_{GS} and WM_{GS}^{200}, even though the scores of FDDVARS of the 370 scans deviated more from zero.

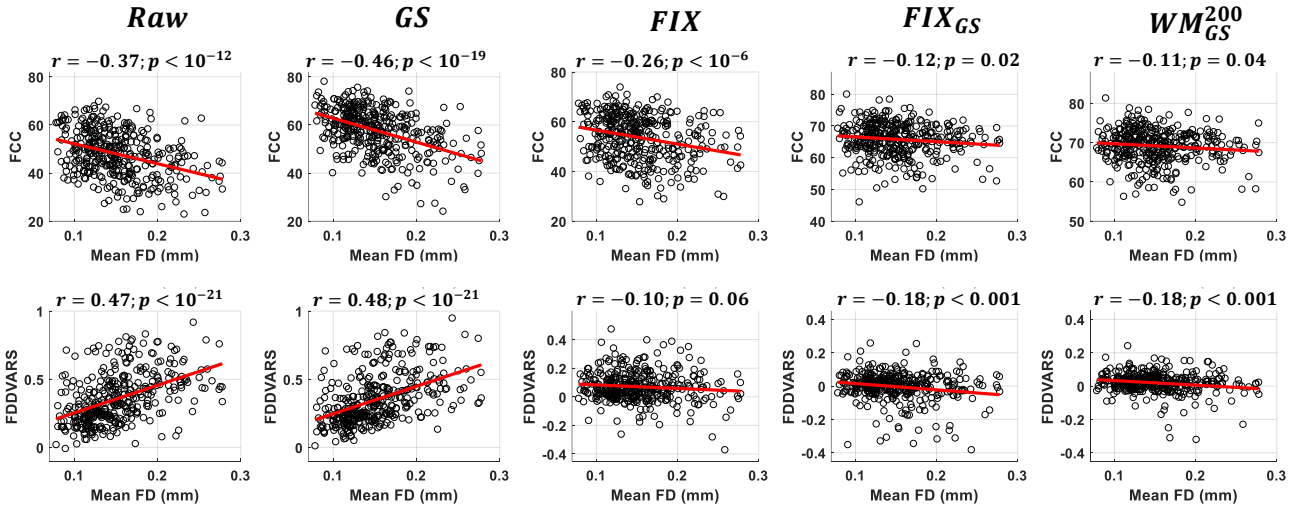


Fig. 6. Scatterplots of mean FD vs FCC (top) and mean FD vs FDDVARS (bottom) considering the first scan from all subjects*. In raw data, the higher were the levels of motion in a scan the more difficult was to identify the networks (low FCC) and the stronger were the motion artifacts in the fMRI data (high FDDVARS). Using the pipelines WM_{GS}^{200} and FIX, the dependence of FCC and FDDVARS on the levels of motions was significantly reduced. *Scans with mean FD three median absolute deviations (MADs) above the median were excluded (based on this criterion, 20 subjects were excluded).

$FDFC_{median}$

When GS was included in the sets of regressors, the scores for $FDFC_{median}$ exhibited a monotonically decreasing trend for varying number of components, starting from 0.15 for raw data (z-score: 0.1) and reaching to 0.13 (z-score: 0.7) for both FIX_{GS} and WM_{GS}^{600} (Fig. 2g). However, when GS was not included in the preprocessing, increasing the number of WM components from 1 to 7 PCA regressors, resulted to an increase of the $FDFC_{median}$ from 0.15 to 0.22 and for higher number of components $FDFC_{median}$ started decreasing reaching again 0.13 with WM_{GS}^{600}.

$FDFC_{dist}$

In the raw data, $FDFC_{dist}$ was -0.16 which reflects that the closer was a parcel to another one the higher was the inflation in their pairwise correlation due to motion (Fig. 2i). Increasing the number of components in WM denoising, resulted in a decrease of the correlation with the more aggressive sets WM^{600} and WM_{GS}^{600} achieving the minimum $FDFC_{dist}$ scores of -0.01 and -0.04. However, the associated z-scores for the latter two sets were relatively low (1.1 and 1.0; Suppl. Fig. 1i) and, as a consequence, $FDFC_{dist}$ did not have significant weighting on the CQC metric.

FD-MFC

FD-MFC was proposed in this work and is based on the assumption that the more a subject moves during a scan the higher is the mean value of correlations in the FC matrix averaged across all edges. As we see in Fig. 2h, the score for FD-MFC in raw data was 0.22 confirming that motion can inflate the estimated correlations in the FC matrix. Importantly, when GSR was not performed, increasing the number of WM components from 1 to 30, led to an increase of FD-MFC with WM^{30} exhibiting an FD-MFC score of 0.45. For higher number of WM components, FD-MFC decreased monotonically reaching 0.19 with WM^{600}. Overall, when WM denoising was combined with GSR demonstrated lower FD-MFC with WM_{GS}^{200} yielding a score of -0.06 (z-score: 0.7). Similar results were found when FD-MFC was estimated using the first scan from all subjects, even though there was a somewhat decrease in the scores for all pipelines (Fig. 7).

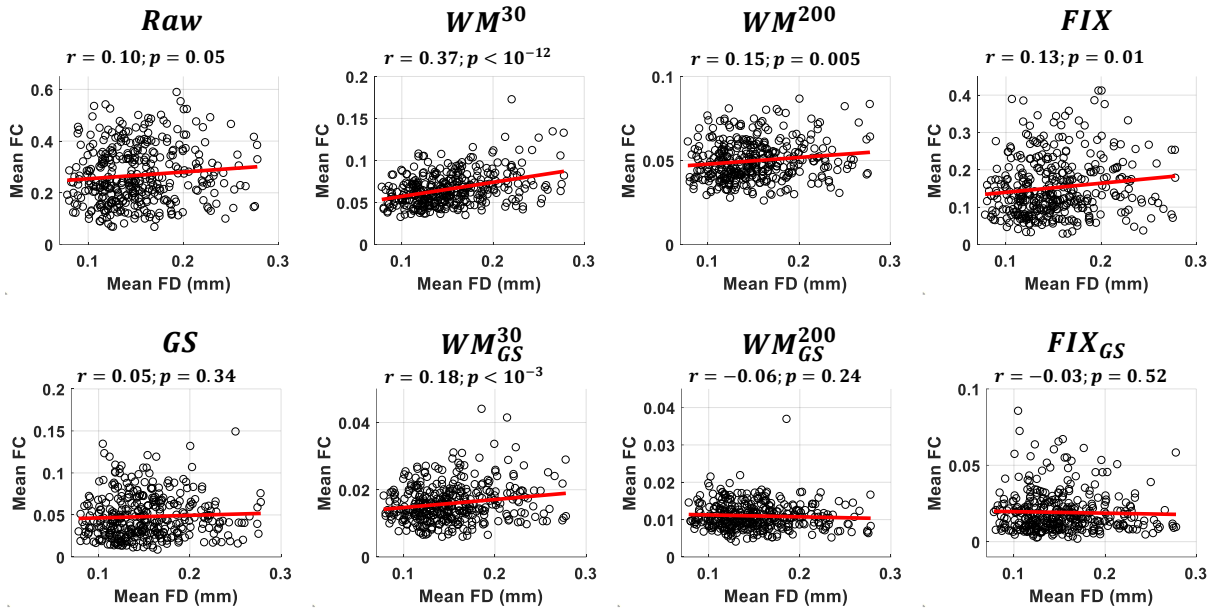


Fig. 7. Scatterplots of mean FD vs mean FC for different pipelines with (bottom) or without (top) GSR considering the first scan from all subjects*. In raw data, the higher were the levels of motion within a scan the higher were the estimated correlations in FC. This dependence on the levels of motion was vanished when the data were preprocessed with WM_{GS}^{200} or FIX_{GS}. Importantly, when a relatively low number of components were removed (e.g., WM^{30}), the effect of motion was enhanced compared to the raw data. *Scans with mean FD three median absolute deviations (MADs) above the median were excluded (based on this criterion, 20 subjects were excluded).

3.2 Evaluation of data-driven NCTs employed in previous studies

In this analysis, we used the QC metrics to compare twenty different pipelines involving the removal of data-driven nuisance regressors from the fMRI data (Table 1). Fig. 8 shows the scores for the summarized metrics QC_{signal} and QC_{motion}, as well as the combined metric CQC. Looking at the first three pipelines that correspond to the 6, 12 and 24 MPs, we observe that motion regressors reduced the effect of motion and to a less extent improved the SNR in the data, with the more aggressive pipeline (24 MPs) exhibiting the strongest impact for all three atlases. GSR alone (pipeline 4) significantly improved the SNR even though, for the Seitzman and MIST atlas, it also led to a small decrease in the QC_{motion} score. As can be seen from Suppl. Fig. 2-Suppl. Fig. 3, FD-FCC and FD-MFC were increased with GSR while FDDVARS was at similar or lower levels compared to the raw data suggesting that even though there was not any enhancement of motion artifacts rather than decrease in the case of the Seitzman atlas, the systematic differences across scans due to motion were increased.

Several studies employ aCompCor as NCT removing five WM and five CSF regressors (Wang et al., 2017; Xiao et al., 2016). Our results derived from the HCP data suggest that this set of regressors demonstrates a moderate improvement with respect to both QC_{signal} and QC_{motion} (pipeline 6). Similar improvement in the quality of data was achieved when the mean time series from WM and CSF, and the 12 MP were regressed out (pipeline 7; Urchs et al., 2017) whereas when including also the GS to the set of regressors the QC_{signal} score reached higher value (pipeline 8; Finn et al., 2015). Pipelines 9 and 10 were more aggressive variants of pipeline 8 that included 24 instead of 12 MPs, as well as the derivatives and squared terms of the tissue mean timeseries from GM, WM and CSF (Ciric et al., 2016; Laumann et al., 2017; Xia et al., 2018b). Considering more nuisance regressors in the preprocessing (36 rather than 15 regressors) pipeline 10 exhibited a small but significant improvement compared to pipeline 8, in both QC_{signal} and QC_{motion} scores,

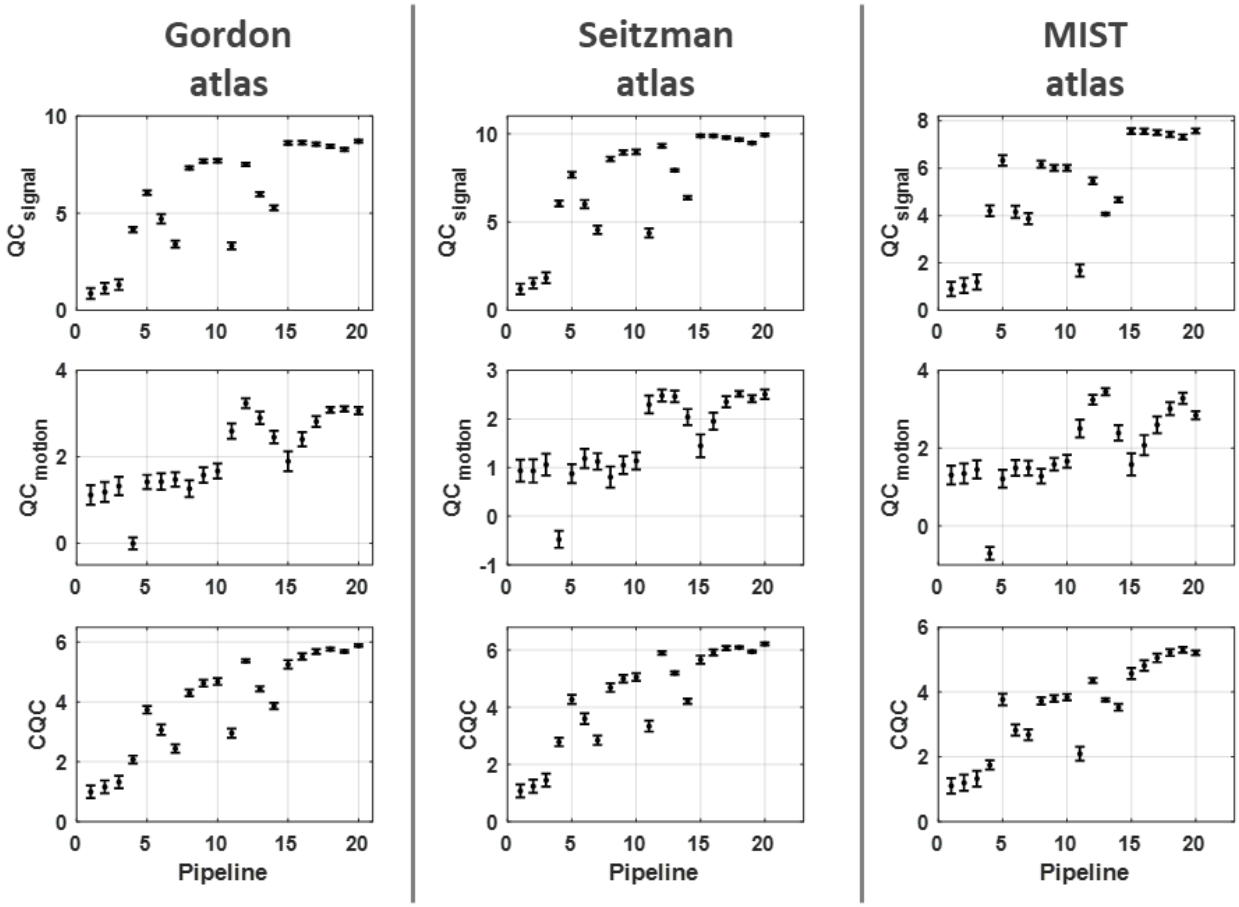


Fig. 8. Evaluation of data-driven NCTs. Twenty different data-driven pipelines were examined listed in Table 1. Among all pipelines, pipelines that consisted of GSR and WM or FIX denoising yielded the highest scores in QC_{signal}, QC_{motion} and CQC (i.e., pipelines 13 and 18-20).

for the Gordon and Seitzman atlases, whereas for the MIST atlas while the QC_{motion} score increased QC_{signal} exhibited a small drop.

Pipelines 11 to 13 evaluated the data quality for the FIX-denoised data provided in HCP with and without further denoising (Fig. 8). We observe that, as proposed in (Burgess et al., 2016), regressing out the GS from the FIX-denoised data improved both QC_{signal} and QC_{motion} scores (pipelines 11 vs 12). However, when five WM and five CSF regressors were removed in addition to the GS (pipeline 13; Siegel et al., 2017) both summarized metrics were lower compared to performing only GSR (pipeline 12).

Pipeline 14 was based on the NCT recommended by Muschelli et al. (2014) which considers as set of regressors the necessary number of WM and CSF regressors needed to explain 50% of variance in their associated compartments. As we see, in all three atlases pipeline 14 exhibited fairly good reduction in motion artifacts even though the SNR was much lower compared to other pipelines. Earlier results presented here showed that, based on the QC_{signal} metric, SNR was relatively low when CSF denoising was performed but high in WM denoising, and particularly when GSR was also performed (Fig. 3). Based on these results, we also considered pipelines 15 to 19 that consider the GS as well as the WM regressors needed to explain a predefined fraction of variance in WM ranging from 30 to 50%. Our results suggest that pipelines 15 to 19 exhibited high scores for both QC_{signal} and QC_{motion} metrics with the highest scores achieved when 45-50% of the variance was used as a threshold to select the WM regressors.

Finally, the set of regressors WM_{GS}^{200} that was found in the previous section to perform the best was considered as pipeline 20. Overall pipelines, we observe that the highest QC scores were obtained when GSR was performed in combination with FIX or WM denoising (i.e., pipelines 13 and 18-20).

3.3 Evaluation of model-based (motion and physiological) NCTs

Four sets of model-based regressors were examined with respect to improvement in SNR and reduction of motion artifacts and biases. The four sets were related to head motion (24 MPs), cardiac pulsatility (modeled with 3rd order RETROICOR), breathing motion (modelled with 3rd order RETROICOR) and SLFOs (modelled with scan-specific physiological response functions with a method proposed in our previous study (Kassinopoulos and Mitsis, 2019; for more information on how the model-based regressors were obtained, see Section 2.4). To assess their contribution when tissue-based regressors are also included in the set of nuisance regressors, we examined 64 pipelines presented in Fig. 9 in the form of a design matrix that refer to combinations of the four sets of model-based regressors, the GS and a set of 200 PCA regressors from WM.

As we can see, when only model-based regressors were considered, accounting for SLFOs improved the QC_{signal} score, whereas correcting for either head or breathing motion improved both QC_{signal} and QC_{motion} scores. Accounting for cardiac pulsatility led to an increase in QC_{signal} and decrease in QC_{motion}, even though the effect of cardiac regressors was lower compared to the rest of the model-based regressors. Finally, when GS and 200 WM regressors were considered (WM_{GS}^{200}), accounting also for breathing motion, cardiac pulsatility or SLFOs, using model-based regressors, did not have any impact on the data quality whereas, in contrary, correcting for motion with the 24 MPs led to a small decrease in the score for QC_{motion}.

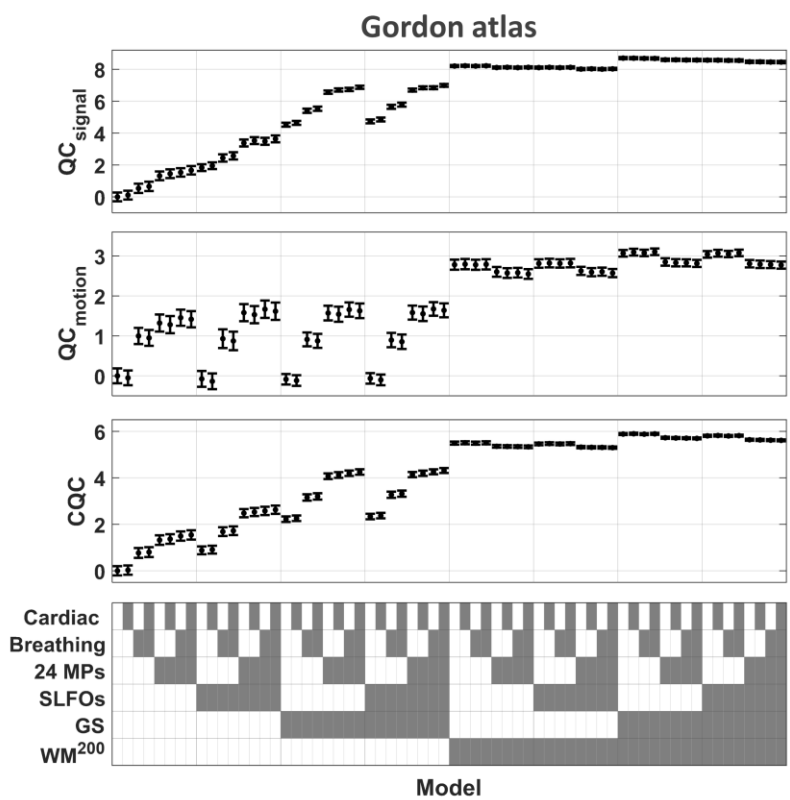


Fig. 9. Evaluation of model-based NCTs using the fMRI data in the Gordon parcel space. Model-based regressors were obtained from the motion (realignment) parameters and physiological recordings to correct for artifacts due to head motion (24 MPs), cardiac pulsatility (Cardiac), breathing motion (Breathing) and SLFOs (i.e., BOLD fluctuations due to changes in heart rate and respiratory flow; Kassinopoulos and Mitsis, 2019). Overall, none of the examined model-based NCTs contributed further to the data quality beyond the improvement achieved with the set of tissue-based regressors WM_{GS}^{200}. The results shown here were obtained from the data in the Gordon parcel space. Similar results were found using the data in the Seitzman and MIST parcel space (Suppl. Fig. 8).

3.4 Evaluation of scrubbing

Discarding volumes contaminated with motion artifacts before regressing out the set of nuisance regressors WM_{GS}^{200} did not provide any gain with respect to the fMRI data quality (Fig. 10). More precisely, the stricter was the thresholds FD_{thr} the lower were the QC_{signal} and QC_{motion}. Also, discarding volumes with DVARS values beyond the threshold did not have any impact on the QC_{motion} score while it significantly decreased the QC_{signal}. Similar results were obtained for the data registered at the Seitzman and MIST atlas (Suppl. Fig. 9).

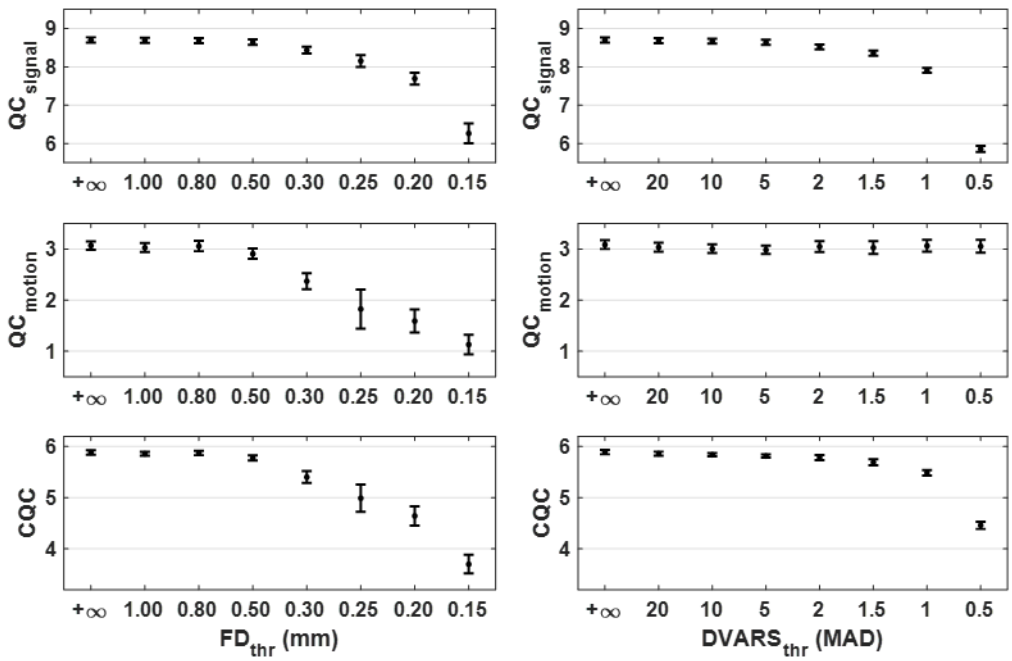


Fig. 10. Effect of scrubbing in data quality for different threshold values. The framewise data quality indices FD and DVARS were used to flag volumes contaminated with motion artifacts. Subsequently, the motion-contaminated volumes were discarded before preprocessing the data with the set WM_{GS}^{200} and estimating the QC_{signal}, QC_{motion} and CQC scores. The obtained scores for varying values of thresholds FD_{thr} and $DVARS_{thr}$ are shown on the left and right columns, respectively. For both FD and DVARS scrubbing, the lower (stricter) were the threshold values the worse was the data quality. Similar results were found using the data in the Seitzman and MIST parcel space (Suppl. Fig. 9).

3.5 Evaluation of low-pass filtering

Considering that the signal of interest in resting-state FC is typically in the low frequencies (<0.10 Hz), a NCT typically used is to low-pass filter the data and the nuisance regressors before their removal so that high-frequency fluctuations attributed to non-neural sources are discarded from the data. To examine the effect of low-pass filtering on data quality as well as its dependence on the cut-off frequency we repeated the denoising of the data with linear regression of the set WM_{GS}^{200} after low-pass filtering the data and the regressors for different cut-off frequencies. As we see in Fig. 11, the highest CQC score was achieved when low-pass filtering was done at 0.20 Hz. Specifically, at this frequency the QC_{signal} was found to be increased by 5% compared to the data that had not been preprocessed (denoted on Fig. 11 with a ∞ cut-off frequency) while the QC_{motion} was kept at similar levels. Importantly, at a 0.08 Hz cut-off frequency that is commonly used in the literature, both the QC_{signal} and QC_{motion} scores decreased by 6% compared to the unfiltered data. Regarding the data registered at the Seitzman and MIST atlases, even though the QC_{signal} and QC_{motion} scores exhibited slightly different trends compared to the Gordon atlas, the cut-off frequency 0.20 Hz achieved again the highest CQC score while the cut-off frequency 0.08 Hz, in the case of the MIST atlas, yielded significantly lower CQC values compared to the unfiltered data (Suppl. Fig. 10).

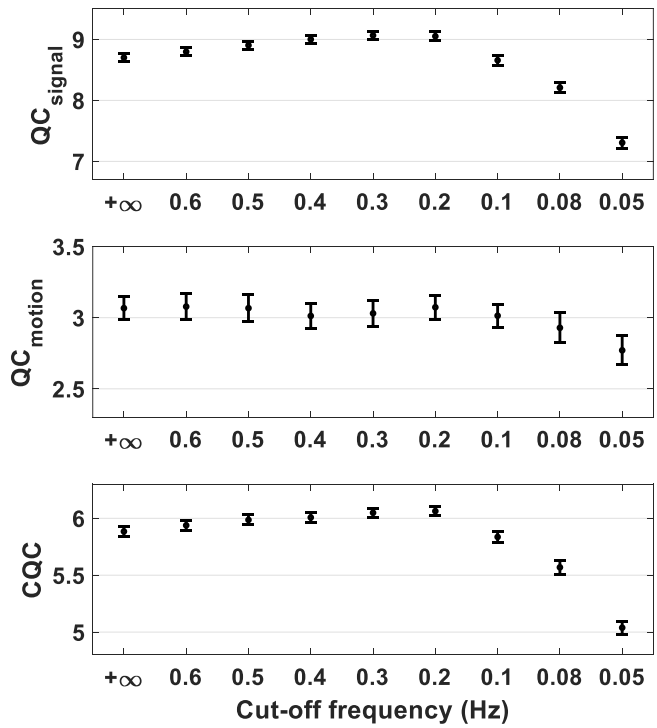


Fig. 11. Effect of low-pass filtering in data quality for different cut-off frequencies. Among all frequencies, low-pass filtering with a cut-off frequency of 0.2 Hz exhibited the highest CQC score. At this cut-off frequency, the QC_{signal} was found to be increased by 5% compared to the unfiltered data, denoted with a ∞ cut-off frequency. The cut-off frequency of 0.2 Hz yielded the highest CQC score also for the data registered in the Seitzman and MIST atlases (Suppl. Fig. 10).

3.6 Identifiability of large-scale networks

Finally, we sought to quantify the identifiability of each of the large-scale networks defined in the three functional atlases employed here and their dependence on the preprocessing pipeline. To this end, we calculated the FCC score of each network for the raw dataset as well as four preprocessed datasets. To obtain the FCC score per network, for a given network, when estimating the FCC score, we compared WNEs with BNEs considering only WNEs belonging to the examined network (for more information see Section 2.7).

In Fig. 12 we see that for the Gordon and Seitzman atlas there was larger variability in FCC score across networks rather than across pipelines. Networks consisting of a small number of parcels, such as the salience network in the Gordon atlas and the medial temporal lobe network in Seitzman atlas, exhibited small negative FCC scores for the raw data whereas when the data were preprocessed with a pipeline that included GSR the FCC scores were increased to small positive values. On the other hand, large networks such as the default mode network exhibited significantly higher FCC scores.

In the case of the MIST atlas there was less variability in FCC score across networks compared to the Gordon and Seitzman atlas which may be because these networks consisted of similar numbers of parcels. That said, two out of the seven networks demonstrated a somewhat strange behavior. Specifically, the mesolimbic network demonstrated a large negative FCC score for the raw and FIX-denoised data despite the fact that it consists of a similar number of parcels as other networks in the atlas. Furthermore, regarding the cerebellum network, even though the FCC score in the raw data was relatively high, when FIX denoising was applied the FCC score dropped to zero.

Finally, while some networks in the three atlases were assigned the same name, did not demonstrate the same behavior in terms of differences in FCC across the five fMRI datasets. For example, in the Gordon atlas we observe that the fronto-pariental network yielded the highest FCC score when the data were preprocessed with the set WM_{GS}^{200} whereas in the MIST atlas the raw data yielded the highest FCC score. Nevertheless, for the majority of networks, FCC scores were maximized when preprocessing was done with FIX_{GS} or WM_{GS}^{200}.

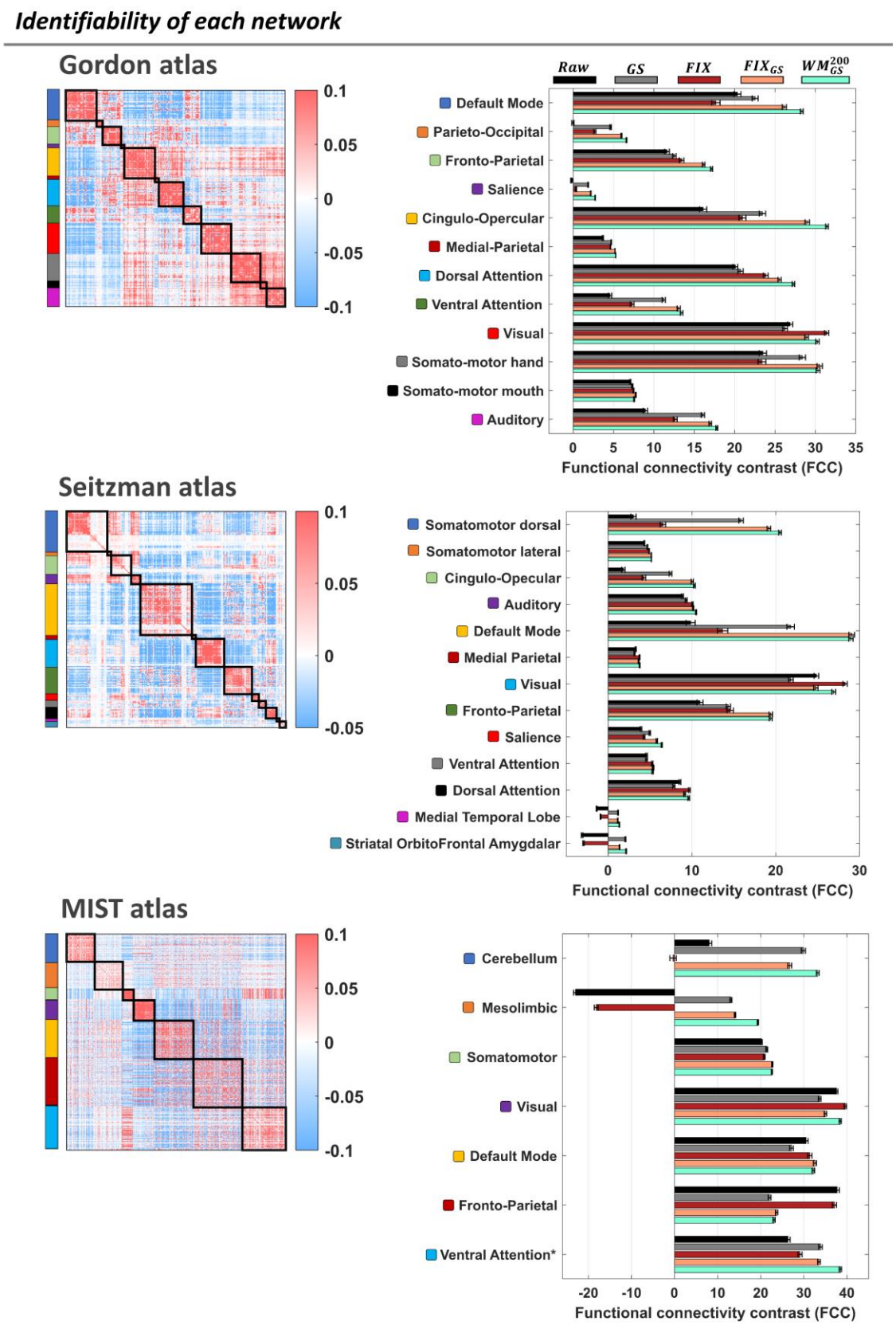


Fig. 12. Identifiability of each network of the three functional atlases Gordon, Seitzman and MIST. The FCC score of each network was defined as the z-statistic of the Wilcoxon rank-sum test related to the null hypothesis that WNEs of the examined network and BNEs in the FC matrix are samples from continuous distributions with equal medians (for more information see Section 2.7). In the case of the Gordon and Seitzman atlases, there is larger variability in FCC scores across networks rather than across pipelines which might be due to the variability in the number of parcels that each network consists of. Note that in the majority of networks, pipelines FIX_{GS} and WM_{GS}^{200} exhibited the highest FCC scores. * The last network in the MIST atlas, apart from the ventral attention network, consists also of the salience network, the basal ganglia and the thalamus.

4. Discussion

In this study, we have rigorously examined the effect of different preprocessing steps on SNR and degree of motion artifacts and biases in resting-state fMRI data. As in previous studies, the QC metrics used to compare preprocessing pipelines illustrated different trends between them (Fig. 2). Therefore, to ease the comparison across pipelines, we introduced a new framework that first normalizes each of the 8 QC metrics to z-scores so that they reflect relative improvement in standard deviations with respect to the raw data. Subsequently, the two normalized signal-related metrics FCC and ICCC and the six normalized motion-related metrics FD-FCC, FDDVARS, FD-FDDVARS, $FDFC_{median}$, $FDFC_{dist}$ and FD-MFC are averaged to obtain the metrics QC_{signal} and QC_{motion}, respectively. Finally, the combined QC metric CQC defined as the mean of the QC_{signal} and QC_{motion} scores is calculated. Using this framework and resting-state fMRI data from the HCP registered to the Gordon atlas, we found that the best data quality was obtained when the GS and 200 PCA regressors from WM were regressed out (Fig. 3). Similar results were found with the fMRI data registered to the Seitzman and MIST atlases as well (Suppl. Fig. 2-Suppl**.** Fig. 5). Note that 200 WM regressors correspond to about 17% of the regressors derived with PCA from WM as the fMRI scans consisted of 1160 volumes each, and explain on average 36±6% of the variance in WM voxel time series.

Despite the fact that we considered in the study only subjects with good quality physiological data in all four scans, none of the model-based techniques examined here exhibited further improvement in terms of data quality when compared to WM denoising (Fig. 9). This may not be surprisingly as it has been previously shown that artifacts due to head motion and physiological fluctuations can be corrected with aCompCor (i.e., removal of five WM and five CSF regressors) as well (Behzadi et al., 2007; Muschelli et al., 2014). Also, WM denoising, and in general model-free approaches such as FIX (Salimi-Khorshidi et al., 2014) and AROMA (Pruim et al., 2015), have the benefit that they do not require physiological data and are not based on any assumptions imposed in physiological models that are likely to be inaccurate. For example, the convolution models used here to account for the effect of heart rate and breathing pattern assume that a linear stationary system can describe these effects which may not be entirely true (Kassinopoulos and Mitsis, 2019). Bear in mind though that the QC metrics considered here and in previous studies reflect biases related to motion rather than physiological processes. As such, we cannot exclude the possibility that physiological model-based techniques may account for differences in physiological variables such as mean heart rate, and in future studies we will try to examine the aforementioned possibility. Moreover, we acknowledge the importance of collecting physiological data in several cases such as when the effects of autonomic nervous system (ANS) or fluctuations in arousal levels are of interest as both ANS and arousal levels are associated to physiological processes (Bonnet and Arand, 1997; Olbrich et al., 2011).

Performing scrubbing before WM denoising was found to harm the quality of the data rather than improving it (Fig. 10). This finding is consistent with Muschelli et al. (2014), who found no improvement with scrubbing when it was followed by aCompCor. However, more recent studies have reported that scrubbing provided some reduction in the score of the motor-related metric $FDFC_{median}$ (Ciric et al., 2017; Parkes et al., 2018) . While these studies employed more sophisticated techniques to correct for motion-contaminated volumes, milder pipelines were also considered compared to the preprocessing pipeline examined here (WM_{GS}^{200}). As such, the potential value of scrubbing cannot be conclusively determined from our study.

Finally, we found that low-pass filtering at 0.2 Hz led to some further improvement in data quality beyond the improvement achieved with WM denoising (Fig. 11). However, substantial decrease in SNR was observed when the 0.08 Hz cut-off frequency commonly used in fMRI studies was considered. The rationale behind choosing the 0.08 Hz cut-off frequency for low-pass filtering in resting-state FC is that well-established large-scale networks have been found to oscillate at frequencies below 0.10 Hz (Damoiseaux et al., 2006) while breathing motion and other sources of noise appear at frequencies above this frequency (Caballero-Gaudes and Reynolds, 2017). Nevertheless, several studies have found activity in RSNs in the range from 0.1 to 0.5 Hz (Chen and Glover, 2015; Niazy et al., 2011) suggesting that low-pass filtering at 0.08 Hz may potentially remove signal of interest. Based on our results, low-pass filtering at 0.2 Hz yields the highest SNR considering whole-brain FC which may be related to reduction in breathing motion artifacts that appear at around 0.3 Hz and might not be fully corrected with WM denoising.

4.1 PCA-based WM denoising improves SNR and mitigates motion effects

In the original study introducing the aCompCor technique (Behzadi et al. 2007) the authors proposed the removal of 6 PCA regressors from WM and CSF to account for cardiac and breathing artifacts. However, this statement was based on Monte Carlo simulations using a modified version of the "broken stick" method described in (Jackson, 2016) which does not take into account QC metrics that reflect in some way improvement in the quality of the fMRI data. A few years later, Chai et al., (2012) also proposed the removal of five PCA regressors from each noise ROI based on observations related to the connectivity of a region in the medial prefrontal cortex with other brain regions. They also showed that regressing out higher number of PCA regressors led to reduced correlation strengths which may be associated to reduction of degrees of freedom in the data. Very likely based on these findings many subsequent fMRI studies considered only 5 PCA regressors from each noise ROI (Ciric et al., 2017; Wang et al., 2017; Xiao et al., 2016).

In this study, we sought to examine the effect of varying the number of PCA regressors on data quality based on QC metrics that account for the effect of motion as well as the SNR in whole-brain FC rather than to interactions between specific regions. Moreover, as there is evidence that neuronal-related activation can be detected in WM (Grajauskas et al., 2019), we examined separately the effect of WM and CSF denoising to determine whether CSF denoising could be sufficient for preprocessing. Interestingly, our results showed that even though WM and CSF denoising achieved similar reduction in motor artifacts and biases, the former exhibited substantially better improvement in SNR than the latter (Fig. 2-Fig. 3). Particularly, the set of regressors WM_{GS}^{200} which consists of 200 PCA regressors from WM and the GS illustrated the best overall performance from all sets of nuisance regressors examined here (Fig. 8).

The standard aCompCor technique that employs 5 PCA regressors from each noise ROI was found to increase the summarized metrics QC_{signal} and QC_{motion} compared to the raw data but not as much as the set WM_{GS}^{200}(pipeline 6 vs 20 in Fig. 8). However, we observed that, when GSR was not considered, removing low number of WM or CSF components exhibited more negative scores in $FDFC_{median}$ and FD-MFC compared to the raw data suggesting that biases in FC due to differences in motion across scans were enhanced (Fig. 2). While this may seem counterintuitive, a possible explanation that we came up with based on Fig. 7 is that in raw data high-motion scans have stronger inflation in connectivity due to motion artifacts than low-motion scans, and even though the first few PCA regressors correct for this inflation, this is done better for low-motion scans with the result of increasing the differences in inflation even more between low- and high-motion scans. This phenomenon was not observed in the scores of $FDFC_{median}$ when GSR was considered and it was diminished for the case of FD-MFC suggesting that the inflation in connectivity may be associated to motion-related fluctuations reflected in the GS as well.

While the practice of regressing out from the data 200 WM regressors may raise concerns with regards to loss of signal of interest, it is important to bear in mind that the examined fMRI data last about 15 minutes and have a repetition time TR of 0.72. Therefore, each of the scans examined here correspond to the relatively large number of 1200 volumes. As a result, the voxel timeseries in WM and CSF were decomposed into 1200 PCA components (note though that the first 40 volumes were subsequently discarded to allow modelling of the SLFOs; for more information see Section 2.4). It is very likely that for shorter duration of data or with a longer TR, a lower number of PCA regressors would yield the best performance and vice versa. Note also that during the training phase of FIX conducted by the HCP group, the average number of components estimated by ICA was 229 and from these components, on average 205 components were labelled as noisy (Stephen M. Smith et al., 2013a) which suggests that finding the set WM_{GS}^{200} performing the best may not be unreasonable.

An alternative preprocessing strategy proposed by Muschelli et al. (2014) is to use the number of PCA regressors needed to explain 50% variance in the two noise ROIs . To compare the performance of this strategy, referred to as aCompCor50, with the original aCompCor they used the QC metric FDDVARS as well as two metrics similar to the FD-FDDVARS and FCC used here. Based on their results, aCompCor50 compared to aCompCor exhibited better reduction in motion artifacts and improvement in specificity in FC even though the difference for the latter was only marginal when corrected for multiple comparisons. In our dataset, aCompCor50 also performed better compared to aCompCor (pipelines 14 vs 6 in Fig. 8). Nevertheless, as we observed the SNR with CSF denoising was lower than with WM denoising and that GSR

seemed to increase the SNR (Fig. 2-Fig. 3), we examined variants of aCompCor50 that consisted of GSR and WM denoising with different thresholds of variance for choosing the number of regressors (pipelines 15-19). In the case of the Gordon and Seitzman atlas, GSR combined with WM regressors needed to explain about 45% variance performed almost as good as WM_{GS}^{200} whereas for the MIST atlas, GSR with WM regressors needed to explain 50% variance performed slightly better than WM_{GS}^{200} (Fig. 8).

4.2 GSR combined with WM or FIX denoising further improves SNR and mitigates motion effects

GS defined as the average fMRI time series across all voxels in the brain or GM is often estimated in order to be regressed out from the data. In our study, GSR improved the scores for the signal-related QC metrics and, to a less extent, the scores for the motion-related QC metrics for both WM and FIX denoising (Fig. 2-Fig. 3). Looking at low number of PCA regressors in WM denoising we observe that the effect of GSR was stronger than in high number of PCA regressors which may be partly due to that WM regressors share common variance with the GS. Previous studies have shown that the GS derived either by the whole brain or GM is almost the same and also that the GS is highly correlated with the mean time series across voxels in WM and CSF (Kassinopoulos and Mitsis, 2019; Power et al., 2017) which supports the idea that WM regressors share common variance with the GS. Furthermore, note that the SLFOs that reflect BOLD fluctuations due to changes in heart rate and breathing pattern, and consist a main component in the GS fluctuations (Falahpour et al., 2013; Kassinopoulos and Mitsis, 2019), were well explained using the first 20-30 WM and CSF regressors Fig. 1. This result suggests that the practice of considering PCA regressors from WM or CSF exhibits to some extend similar effects with GSR. As a result, the effect of GS when considering 200 WM regressors (i.e., WM_{GS}^{200} vs WM^{200}) is relatively small (Fig. 2-Fig. 3). In contrast, GSR has a strong effect on FIX denoising which may suggest that the ICA regressors that are removed in FIX denoising do not share variance with the GS. This is not surprisingly, as it has been suggested that spatial ICA used in FIX is mathematically, by design, unable to separate global temporal artifacts from fMRI data (Glasser et al., 2018).

Despite the simplicity of GSR, there has been much debate about its use (Liu et al., 2017; Murphy and Fox, 2017). Even though several studies have shown that a large fraction of the GS is associated to physiological processes such as heart rate and breathing activity (Birn et al., 2006; Chang et al., 2009; Falahpour et al., 2013; Kassinopoulos and Mitsis, 2019; Shmueli et al., 2007; Wise et al., 2004) as well as head motion (Power et al., 2014; Satterthwaite et al., 2013), there is accumulating evidence that GS is also driven by neuronal activity as assessed by intracranial recordings (Schölvinck et al., 2010) and vigilance-related measures (Chang et al., 2016; Falahpour et al., 2018; Wong et al., 2013, 2016). Therefore, while our results are in support of GSR for both WM and FIX denoising we cannot exclude the possibility of removing some neuronal-related fluctuations from the data when the GS is removed.

4.3 QC metrics

Nine QC metrics were initially considered with three metrics related to the SNR in the fMRI data and six metrics related to motion artifacts and biases. To assess the sensitivity of each metric, the subjects were split into 10 groups of 39 subjects each with similar levels of motion across groups as assessed with within-scan mean FD. Subsequently, the QC scores were estimated for each group separately. Based on the fact that the 10 groups of subjects were characterized by similar distributions of mean FD values, we considered that the more sensitive a QC metric is the smaller would be the variability (or standard deviation) of scores across groups. And to give some units in the QC metrics that are easier interpretable, the score for a given metric and group of subjects was expressed as a z-score that reflects the improvement in standard deviations compared to the distribution of values found in the raw data across the ten groups of subjects (for more information see Section 2.8).

Signal-related metrics

Among the three signal-related metrics (i.e., metrics related to the SNR), the FCC demonstrated the highest improvement in z-score for varying number of WM regressors (Suppl. Fig. 1). The FCC is based on the assumption that the strength of correlation for WNEs in FC are on average larger than BNEs. Previous studies have used similar metrics to assess spatial specificity in FC considering though only interactions between specific regions in the brain rather than whole-

brain interactions (Birn et al., 2014; Chai et al., 2012; Muschelli et al., 2014) whereas Shirer et al. (2015) used a metric that compares the correlations of WNEs with correlations between brain regions and regions outside the brain. While we acknowledge that some of the BNEs may correspond to neuronal-related connections, these edges would be the minority. Therefore, we believe considering all BNEs to form the null distribution rather than connections with voxels outside the brain is more appropriate as the relative magnitude of within- vs -between- network edges allows essentially the identification of clusters or networks.

The signal-related metric MICC was used to assess test-retest reliability across the four sessions of each subject in whole-brain FC estimates. However, as in previous studies, the more aggressive a pipeline was the lower was the MICC score which has been interpreted as the metric reflecting subject-specificity due to presence of noise rather than signal of interest (Fig. 2, Birn et al., 2014; Parkes et al., 2018). As MICC scores did not seem to correspond to SNR it was excluded from the rest of the analysis. Interestingly, Birn et al. (2014) reported smaller decreases in ICC for significant connections compared to the remaining connections which was also confirmed in our data (look for example Fig. 5). Therefore, in this work, based on these findings, we proposed a novel metric named ICCC that reflects how much higher are the ICC values in WNEs compared to BNEs. ICCC was found to behave in a similar manner with FCC and, thus, later in the analysis was combined with the FCC score to obtain the summarized metric QC_{signal}. Note that when the data were preprocessed with FIX_{GS} or WM_{GS}^{200}, edges corresponding to interactions between the default mode and fronto-parietal network despite the low correlation values in group-level FC, they demonstrated significantly higher ICC values compared to other BNEs (Fig. 5). This finding suggests that regions in the default mode and fronto-parietal networks may be functionally connected but in a subject-specific manner. On a side note, the values of connectivity strength between regions in the aforementioned two networks were found in recent studies to contribute in the identification of individuals using fMRI FC (Finn et al., 2015) as well as in the prediction of behavioral measures (Smith et al., 2015).

A caveat of using ICCC as a metric to compare pipelines is that it requires a dataset with several subjects and more than one scan per subject. As a result, in contrast to FCC, it cannot be used to assess the data quality for a specific scan. In addition, looking at Fig. 2 & Fig. 5, we see that ICCC was increased both with a better preprocessing strategy or with a larger sample size. However, when ICCC was estimated from all 390 subjects in one step rather than in groups of 39 subjects, apart from the increase in ICCC scores for all pipelines we also observe smaller differences between pipelines which can be translated to lower sensitivity of ICCC when comparing pipelines. We found the dependence of the metric ICCC on sample size somewhat puzzling. However, for future studies with large sample sizes interested in assessing the performance of pipelines, we would recommend estimating ICCC in small groups of subjects as done here.

Motion-related metrics

Head motion during the scan is a major confound in fMRI FC studies as it diminishes the signal of interest in the data but also affects the strength of connectivity across regions and across populations in a systematic manner. While the majority of edges in FC are typically inflated by motion, short-distance edges tend to be inflated even more than long-distance edges (Satterthwaite et al., 2013). In addition, different populations present often different tendency for motion (e.g., young vs older participants) which has been shown to lead to artificial differences in FC (Power et al., 2015). To assess the performance of each preprocessing strategy examined here on the aforementioned aspects of motion effects, three previously proposed metrics (i.e., FDDVARS, $FDFC_{median}$ and $FDFC_{dist}$) as well as three new metrics (i.e., FD-FCC, FD-FDDVARS and FD-MFC) were in considered in this study. While the main trend in all motion-related metrics was that the more WM regressors were removed the closer were the scores to zero, a different pipeline was favored from each metric (Fig. 2). For example, considering WM denoising with GSR, the metric FD-FCC demonstrated the smallest absolute score when 70 WM regressors were removed whereas the metric $FDFC_{median}$ yielded the smallest absolute score for the most aggressive pipeline examined here that consisted of 600 WM regressors. However, after normalizing the metrics to z-scores, FDDVARS was found to be considerably more sensitive than the remaining metrics (Suppl. Fig. 1). As a result, the summarized metric QC_{motion} that was defined as the average of all six motion-related metrics, favored the set WM_{GS}^{200} as well.

Combined QC metric

While the summarized metric associated to SNR QC_{signal} reached a maximum score for the broad range of pipelines WM_{GS}^{60} to WM_{GS}^{200}, to ensure an efficient mitigation of motion artifacts and biases, the more aggressive option of WM denoising (i.e., WM_{GS}^{200}) was favored by the combined QC metric (Fig. 3). However, we acknowledge that depending on the fMRI study, the researchers may give more weighting to the metric QC_{signal} and, thus, apply a milder WM denoising, particularly when two populations with similar levels of motion are compared.

5. Conclusion

In summary, the current study evaluated the performance of a large range of data-driven and model-based techniques using previously proposed QC metrics as well as novel metrics. As the QC metrics did not uniformly favor a specific preprocessing strategy, we proposed a framework that evaluates the sensitivity of each metric. Among eight QC metrics, the metric FCC proposed here as well as the metric FDDVARS employed in Muschelli et al. (2014) exhibited the highest sensitivity. FCC reflects how much higher are the correlation values in WNEs compared to BNEs in FC whereas FDDVARS reflects the levels of motion artifacts in the parcel time series. The data-driven approaches WM denoising and FIX denoising combined with GSR demonstrated the largest increase in SNR as well as reduction in motion artifacts and biases. In the case of WM denoising, using resting-state fMRI data from the HCP, we found that about 17% of the WM regressors had to be removed to improve the QC scores. Scrubbing did not provide any gain to the data quality when it was followed by WM denoising, and low-pass filtering at 0.2 Hz increased slightly the SNR.

Similar conclusions were derived using three different functional atlases. However, unless the framework followed here is repeated with different datasets that vary in terms of population examined or acquisition parameters (e.g. repetition time TR and duration of scan) we cannot be certain whether the conclusions derived here can be generalized to other datasets. Therefore, we recommend investigators to consult the QC metrics when deciding about the pipeline they want to employ in a study. Finally, as has been suggested in previous studies (Ciric et al., 2017; Parkes et al., 2018), we recommend investigators to report scores of QC metrics for the preprocessed data so that readers can independently interpret the findings with respect to possible biases that can arise due to motion. To assist with this, we provide the codes used in this study (https://github.com/mkassinopoulos/Estimation_of_QC_metrics) that can be used for preprocessing of the data and estimation of the QC scores.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant 34362 awarded to GDM), the Fonds de la Recherche du Quebec - Nature et Technologies (FRQNT; Team Grant PR191780-2016 awarded to GDM) and the Canada First Research Excellence Fund (awarded to McGill University for the Healthy Brains for Healthy Lives initiative). MK acknowledges funding from Québec Bio-imaging Network (QBIN). Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

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. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bijsterbosch, J., Harrison, S., Duff, E., Alfaro-Almagro, F., Woolrich, M., Smith, S., 2017. Investigations into within- and between-subject resting-state amplitude variations. Neuroimage 159, 57–69. https://doi.org/10.1016/j.neuroimage.2017.07.014

Bijsterbosch, J.D., Woolrich, M.W., Glasser, M.F., Robinson, E.C., Beckmann, C.F., Van Essen, D.C., Harrison, S.J., Smith, S.M., 2018. The relationship between spatial configuration and functional connectivity of brain regions. Elife 7, 210195. https://doi.org/10.7554/eLife.32992

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 Connect. 4, 511–522. https://doi.org/10.1089/brain.2014.0284

Birn, R.M., Diamond, J.B., Smith, M.A., Bandettini, P.A., 2006. Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. Neuroimage 31, 1536–48. https://doi.org/10.1016/j.neuroimage.2006.02.048

Birn, R.M., Smith, M. a., Jones, T.B., Bandettini, P. a., 2008. The respiration response function: The temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage 40, 644–654. https://doi.org/10.1016/j.neuroimage.2007.11.059

Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn. Reson. Med. 34, 537–541. https://doi.org/10.1002/mrm.1910340409

Bonnet, M.H., Arand, D.L., 1997. Heart rate variability: Sleep stage, time of night, and arousal influences. Electroencephalogr. Clin. Neurophysiol. 102, 390–396. https://doi.org/10.1016/S0921-884X(96)96070-1

Burgess, G.C., Kandala, S., Nolan, D., Laumann, T.O., Power, J.D., Adeyemo, B., Harms, M.P., Petersen, S.E., Barch, D.M., 2016. Evaluation of Denoising Strategies to Address Motion-Correlated Artifacts in Resting-State Functional Magnetic Resonance Imaging Data from the Human Connectome Project. Brain Connect. 6, 669–680. https://doi.org/10.1089/brain.2016.0435

Caballero-Gaudes, C., Reynolds, R.C., 2017. Methods for cleaning the BOLD fMRI signal. Neuroimage 154, 128–149. https://doi.org/10.1016/j.neuroimage.2016.12.018

Chai, X.J., Castañán, A.N., Öngür, D., Whitfield-Gabrieli, S., 2012. Anticorrelations in resting state networks without global signal regression. Neuroimage 59, 1420–1428. https://doi.org/10.1016/j.neuroimage.2011.08.048

Chang, C., Cunningham, J.P., Glover, G.H., 2009. Influence of heart rate on the BOLD signal: The cardiac response function. Neuroimage 44, 857–869. https://doi.org/10.1016/j.neuroimage.2008.09.029

Chang, C., Leopold, D.A., Schölvinck, M.L., Mandelkow, H., Picchioni, D., Liu, X., Ye, F.Q., Turchi, J.N., Duyn, J.H., 2016. Tracking brain arousal fluctuations with fMRI. Proc. Natl. Acad. Sci. 113, 4518–4523. https://doi.org/10.1073/pnas.1520613113

Chen, J.E., Glover, G.H., 2015. BOLD fractional contribution to resting-state functional connectivity above 0.1Hz. Neuroimage 107, 207–218. https://doi.org/10.1016/j.neuroimage.2014.12.012

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., 2016. Benchmarking confound regression strategies for the control of motion artifact in studies of functional connectivity. ArXiv. https://doi.org/10.1016/j.neuroimage.2017.03.020

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., 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 154, 174–187. https://doi.org/10.1016/j.neuroimage.2017.03.020

Dagli, M.S., Ingeholm, J.E., Haxby, J. V, 1999. Localization of cardiac-induced signal change in fMRI. Neuroimage 9, 407–415. https://doi.org/10.1006/nimg.1998.0424

Damoiseaux, J.S., Rombouts, S.A.R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks.

Demirtaş, M., Tornador, C., Falcón, C., López-Solà, M., Hernández-Ribas, R., Pujol, J., Menchón, J.M., Ritter, P., Cardoner, N., Soriano-Mas, C., Deco, G., 2016. Dynamic functional connectivity reveals altered variability in functional connectivity among patients with major depressive disorder. Hum. Brain Mapp. 00. https://doi.org/10.1002/hbm.23215

Falahpour, M., Chang, C., Wong, C.W., Liu, T.T., 2018. Template-based prediction of vigilance fluctuations in resting-state fMRI. Neuroimage 174, 317–327. https://doi.org/10.1016/j.neuroimage.2018.03.012

Falahpour, M., Refai, H., Bodurka, J., 2013. Subject specific BOLD fMRI respiratory and cardiac response functions obtained from global signal. Neuroimage 72, 252–264. https://doi.org/10.1016/j.neuroimage.2013.01.050

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671. https://doi.org/10.1038/nn.4135

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc. Natl. Acad. Sci. U. S. A. 102, 9673–8. https://doi.org/10.1073/pnas.0504136102

Glasser, M.F., Coalson, T.S., Bijsterbosch, J.D., Harrison, S.J., Harms, M.P., Anticevic, A., Van Essen, D.C., Smith, S.M., 2018. Using temporal ICA to selectively remove global noise while preserving global signal in functional MRI data. Neuroimage 181, 692–717. https://doi.org/10.1016/j.neuroimage.2018.04.076

Glasser, M.F., Smith, S.M., Marcus, D.S., Andersson, J.L.R., Auerbach, E.J., Behrens, T.E.J., Coalson, T.S., Harms, M.P., Jenkinson, M., Moeller, S., Robinson, E.C., Sotiropoulos, S.N., Xu, J., Yacoub, E., Ugurbil, K., Van Essen, D.C., 2016. The Human Connectome Project's neuroimaging approach. Nat. Neurosci. 19, 1175–87. https://doi.org/10.1038/nn.4361

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Glover, G.H., Li, T.-Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44, 162–167. https://doi.org/10.1002/1522-2594(200007)44:1<162::AID-MRM23>3.0.CO;2-E

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. https://doi.org/10.1093/cercor/bhu239

Grajauskas, L.A., Frizzell, T., Song, X., D'Arcy, R.C.N., 2019. White Matter fMRI Activation Cannot Be Treated as a Nuisance Regressor: Overcoming a Historical Blind Spot. Front. Neurosci. 13, 2007–2010. https://doi.org/10.3389/fnins.2019.01024

Jackson, D.A., 2016. Stopping Rules in Principal Components Analysis : A Comparison of Heuristical and Statistical Approaches Stable URL : http://www.jstor.org/stable/1939574 REFERENCES Linked references are available on JSTOR for this article : You may need to log in to JSTO 74, 2204–2214.

Kassinopoulos, M., Mitsis, G.D., 2019. Identification of physiological response functions to correct for fluctuations in

resting-state fMRI related to heart rate and respiration. Neuroimage 202, 116150. https://doi.org/10.1016/j.neuroimage.2019.116150

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., McCarthy, J.E., Tagliazucchi, E., Laufs, H., Schlaggar, B.L., Dosenbach, N.U.F., Petersen, S.E., 2017. On the Stability of BOLD fMRI Correlations. Cereb. Cortex 27, 4719–4732. https://doi.org/10.1093/cercor/bhw265

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.M., Schluep, M., Vuilleumier, P., Van De Ville, D., 2013. Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. Neuroimage 83, 937–950. https://doi.org/10.1016/j.neuroimage.2013.07.019

Liu, T.T., Nalci, A., Falahpour, M., 2017. The global signal in fMRI: Nuisance or Information? Neuroimage 150, 213–229. https://doi.org/10.1016/j.neuroimage.2017.02.036

Murphy, K., Birn, R.M., Bandettini, P.A., 2013. Resting-state fMRI confounds and cleanup. Neuroimage 80, 349–359. https://doi.org/10.1016/j.neuroimage.2013.04.001

Murphy, K., Fox, M.D., 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. Neuroimage 154, 169–173. https://doi.org/10.1016/j.neuroimage.2016.11.052

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. https://doi.org/10.1016/j.neuroimage.2014.03.028

Niazy, R.K., Xie, J., Miller, K., Beckmann, C.F., Smith, S.M., 2011. Spectral characteristics of resting state networks, 1st ed, Progress in Brain Research. Elsevier B.V. https://doi.org/10.1016/B978-0-444-53839-0.00017-X

Olbrich, S., Sander, C., Matschinger, H., Mergl, R., Trenner, M., Schönknecht, P., Hegerl, U., 2011. Brain and body: Associations between EEG-vigilance and the autonomous nervous system activity during rest. J. Psychophysiol. 25, 190–200. https://doi.org/10.1027/0269-8803/a000061

Parkes, L., Fulcher, B., Yücel, M., Fornito, A., 2018. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage 171, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Perlbarg, V., Bellec, P., Anton, J.L., Pélégrini-Issac, M., Doyon, J., Benali, H., 2007. CORSICA: correction of structured noise in fMRI by automatic identification of ICA components. Magn. Reson. Imaging 25, 35–46. https://doi.org/10.1016/j.mri.2006.09.042

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. https://doi.org/10.1016/j.neuroimage.2011.10.018

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2014.10.044

Prokopiou, P.C., Pattinson, K.T.S., Wise, R.G., Mitsis, G.D., 2019. Modeling of dynamic cerebrovascular reactivity to spontaneous and externally induced CO2 fluctuations in the human brain using BOLD-fMRI. Neuroimage 186, 533–548. https://doi.org/10.1016/j.neuroimage.2018.10.084

Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015. ICA-AROMA: A robust

ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112, 267–277. https://doi.org/10.1016/j.neuroimage.2015.02.064

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. https://doi.org/10.1016/j.neuroimage.2013.11.046

Satterthwaite, T.D., Ciric, R., Roalf, D.R., Davatzikos, C., Bassett, D.S., Wolf, D.H., 2019. Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum. Brain Mapp. 40, 2033–2051. https://doi.org/10.1002/hbm.23665

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. https://doi.org/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. https://doi.org/10.1016/j.neuroimage.2011.12.063

Schölvinck, M.L., Maier, A., Ye, F.Q., Duyn, J.H., Leopold, D.A., 2010. Neural basis of global resting-state fMRI activity. Proc. Natl. Acad. Sci. U. S. A. 107, 10238–43. https://doi.org/10.1073/pnas.0913110107

Seitzman, B.A., Gratton, C., Marek, S., Raut, R. V, Dosenbach, N.U., Schlaggar, B.L., Petersen, S.E., Greene, D.J., 2018. A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. Neuroimage 450452. https://doi.org/10.1101/450452

Shirer, W.R., Jiang, H., Price, C.M., Ng, B., Greicius, M.D., 2015. Optimization of rs-fMRI Pre-processing for Enhanced Signal-Noise Separation, Test-Retest Reliability, and Group Discrimination. Neuroimage 117, 67–79. https://doi.org/10.1016/j.neuroimage.2015.05.015

Shmueli, K., van Gelderen, P., de Zwart, J. a., Horovitz, S.G., Fukunaga, M., Jansma, J.M., Duyn, J.H., 2007. Low-frequency fluctuations in the cardiac rate as a source of variance in the resting-state fMRI BOLD signal. Neuroimage 38, 306–320. https://doi.org/10.1016/j.neuroimage.2007.07.037

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

Siegel, J.S., Mitra, A., Laumann, T.O., Seitzman, B.A., Raichle, M., Corbetta, M., Snyder, A.Z., 2017. Data Quality Influences Observed Links Between Functional Connectivity and Behavior. Cereb. Cortex 27, 4492–4502. https://doi.org/10.1093/cercor/bhw253

Smith, Stephen M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A.T., Woolrich, M.W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D.C., Glasser, M.F., 2013a. Resting-state fMRI in the Human Connectome Project. Neuroimage 80, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

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. Proc. Natl. Acad. Sci. U. S. A. 106, 13040–5. https://doi.org/10.1073/pnas.0905267106

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E.J., Glasser, M.F., Ugurbil, K., Barch, D.M., Van Essen, D.C., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci. 18, 1565–1567. https://doi.org/10.1038/nn.4125

Smith, Stephen M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., Barch, D.M., Uğurbil, K., Van Essen, D.C., 2013b. Functional

connectomics from resting-state fMRI. Trends Cogn. Sci. 17, 666–682. https://doi.org/10.1016/j.tics.2013.09.016

Urchs, S., Armoza, J., Benhajali, Y., St-Aubin, J., Orban, P., Bellec, P., 2017. MIST: A multi-resolution parcellation of functional brain networks. MNI Open Res. 1, 3. https://doi.org/10.12688/mniopenres.12767.1

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. https://doi.org/10.1016/j.neuroimage.2011.07.044

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn Human Connectome Project: An overview. Neuroimage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Wang, J., Ren, Y., Hu, X., Nguyen, V.T., Guo, L., Han, J., Guo, C.C., 2017. Test–retest reliability of functional connectivity networks during naturalistic fMRI paradigms. Hum. Brain Mapp. 38, 2226–2241. https://doi.org/10.1002/hbm.23517

Whittaker, J.R., Driver, I.D., Venzi, M., Bright, M.G., Murphy, K., Chen, J., Whittaker, J.R., 2019. Cerebral Autoregulation Evidenced by Synchronized Low Frequency Oscillations in Blood Pressure and Resting-State fMRI 13, 1–12. https://doi.org/10.3389/fnins.2019.00433

Wise, R.G., Ide, K., Poulin, M.J., Tracey, I., 2004. Resting fluctuations in arterial carbon dioxide induce significant low frequency variations in BOLD signal. Neuroimage 21, 1652–1664. https://doi.org/10.1016/j.neuroimage.2003.11.025

Wong, C.W., DeYoung, P.N., Liu, T.T., 2016. Differences in the resting-state fMRI global signal amplitude between the eyes open and eyes closed states are related to changes in EEG vigilance. Neuroimage 124, 24–31. https://doi.org/10.1016/j.neuroimage.2015.08.053

Wong, C.W., Olafsson, V., Tal, O., Liu, T.T., 2013. The amplitude of the resting-state fMRI global signal is related to EEG vigilance measures. Neuroimage 83, 983–990. https://doi.org/10.1016/j.neuroimage.2013.07.057

Woodward, N.D., Cascio, C.J., 2015. Resting-State Functional Connectivity in Psychiatric Disorders. JAMA psychiatry 72, 743–744. https://doi.org/10.1001/jamapsychiatry.2015.0101.2

Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkins, M.E., Cook, P.A., García de la Garza, A., Vandekar, S.N., Cui, Z., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C., Gur, R.E., Shinohara, R.T., Bassett, D.S., Satterthwaite, T.D., 2018a. Linked dimensions of psychopathology and connectivity in functional brain networks. Nat. Commun. 9, 1–14. https://doi.org/10.1038/s41467-018-05317-y

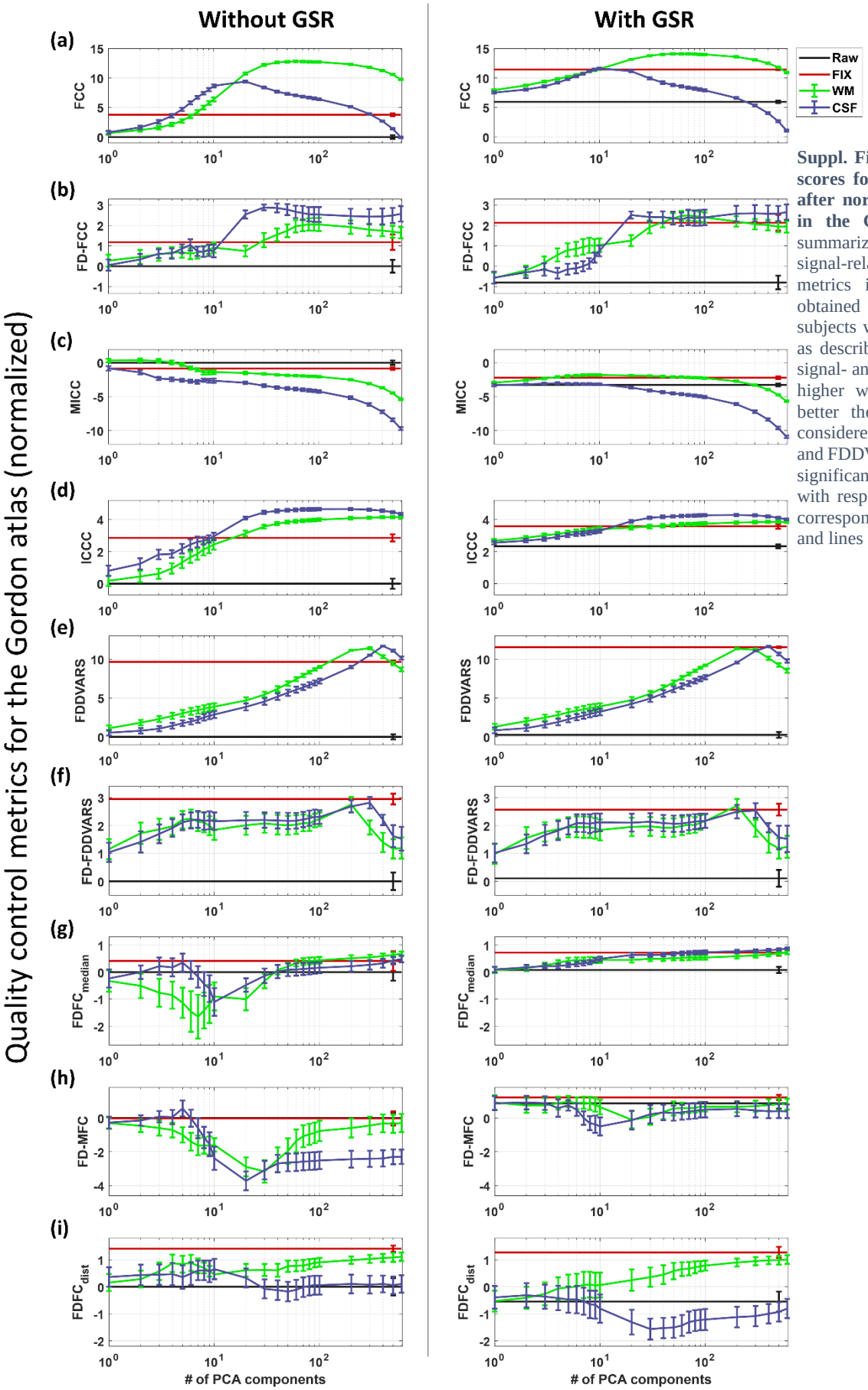
Xia, C.H., Ma, Z., Ciric, R., Gu, S., Betzel, R.F., Kaczkurkin, A.N., Calkins, M.E., Cook, P.A., García de la Garza, A., Vandekar, S.N., Cui, Z., Moore, T.M., Roalf, D.R., Ruparel, K., Wolf, D.H., Davatzikos, C., Gur, R.C.R.E., Gur, R.C.R.E., Shinohara, R.T., Bassett, D.S., Satterthwaite, T.D., Figures, S., Tables, S., 2018b. Linked dimensions of psychopathology and connectivity in functional brain networks. Nat. Commun. 9, 1–14. https://doi.org/10.1038/s41467-018-05317-y

Xiao, Y., Friederici, A.D., Margulies, D.S., Brauer, J., 2016. Longitudinal changes in resting-state fMRI from age 5 to age 6years covary with language development. Neuroimage 128, 116–124. https://doi.org/10.1016/j.neuroimage.2015.12.008

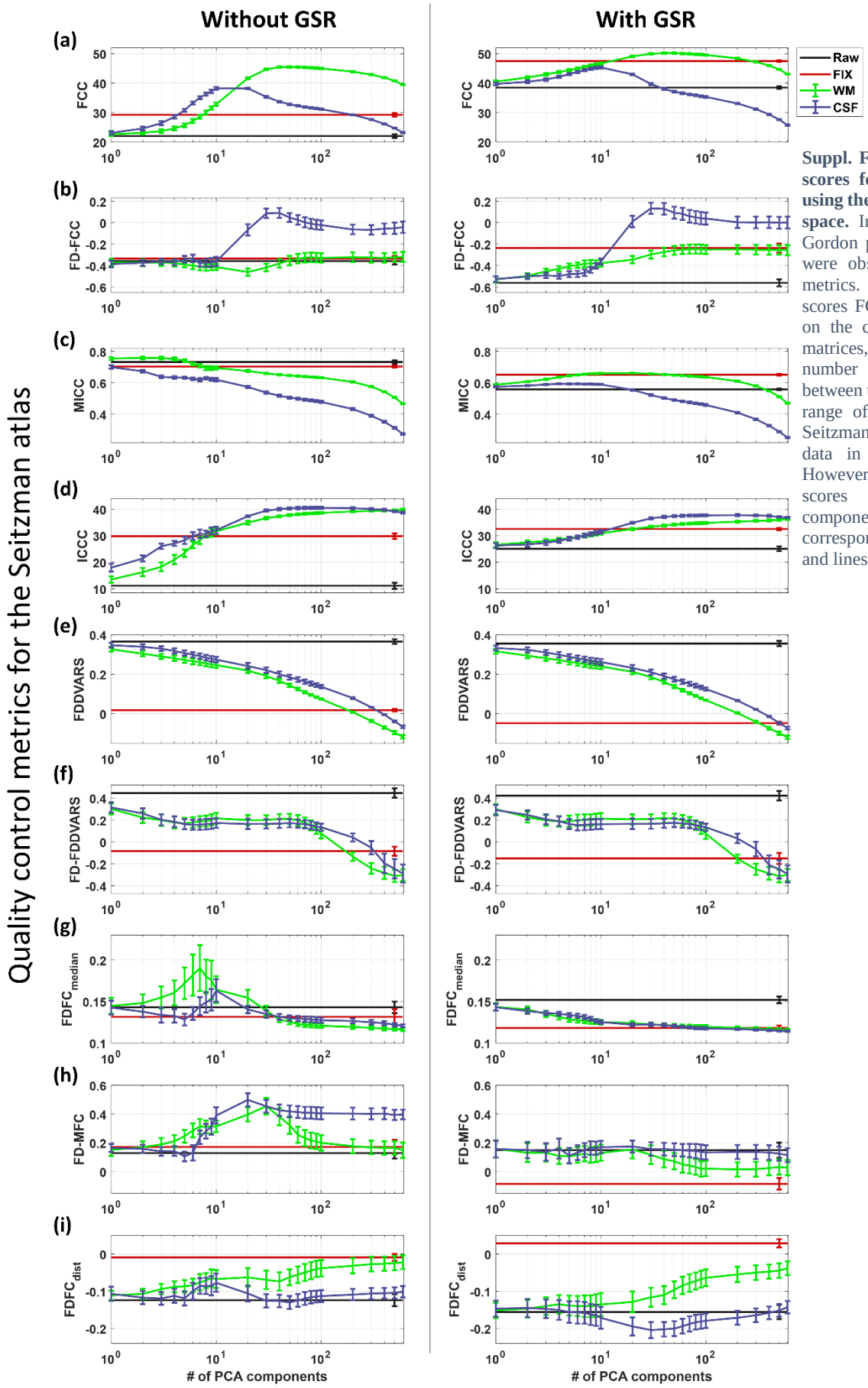
Zhang, C., Dougherty, C.C., Baum, S.A., White, T., Michael, A.M., 2018. Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. Hum. Brain Mapp. 39, 1765–1776. https://doi.org/10.1002/hbm.23950

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57. https://doi.org/10.1109/42.906424

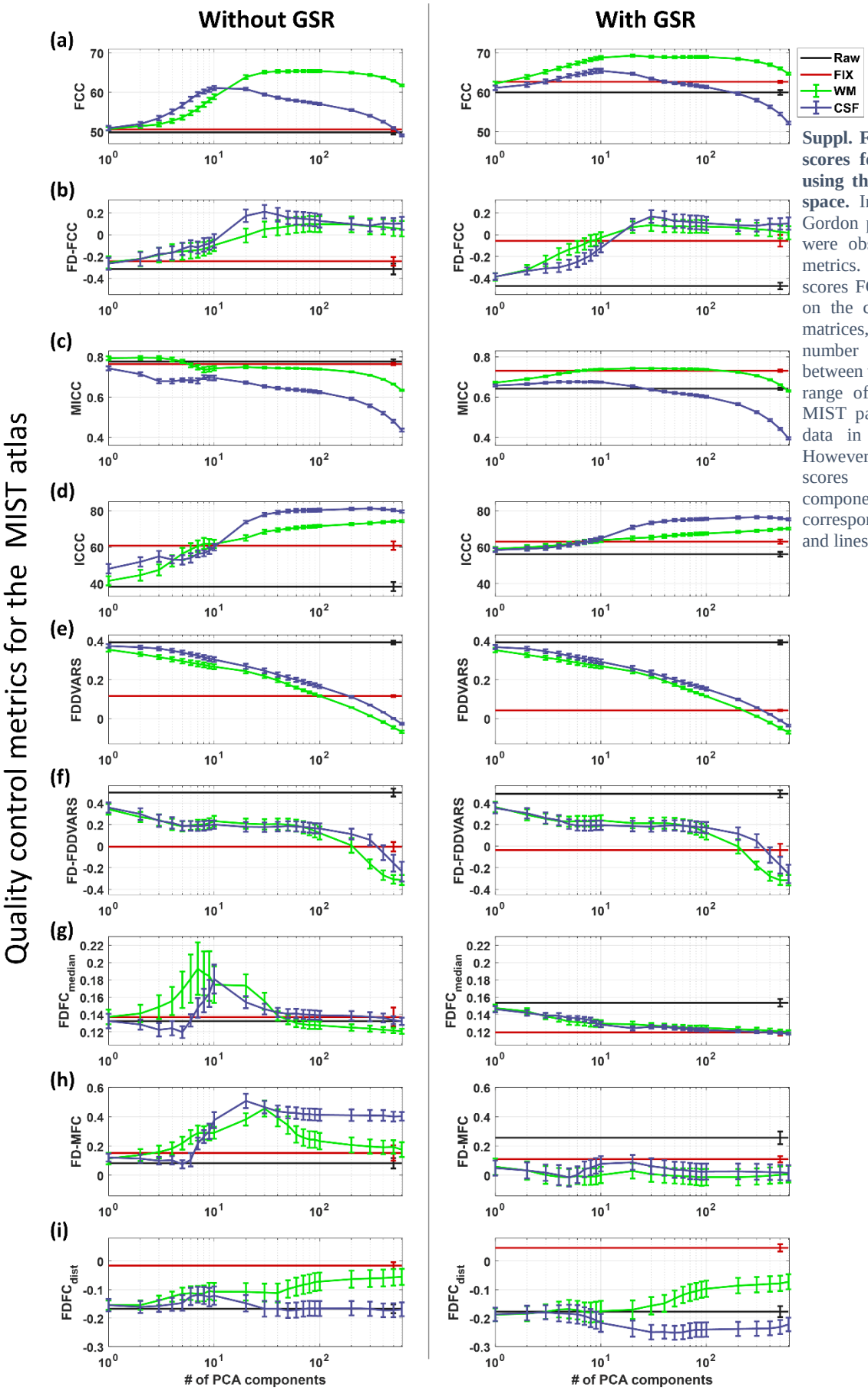
Supplementary Material



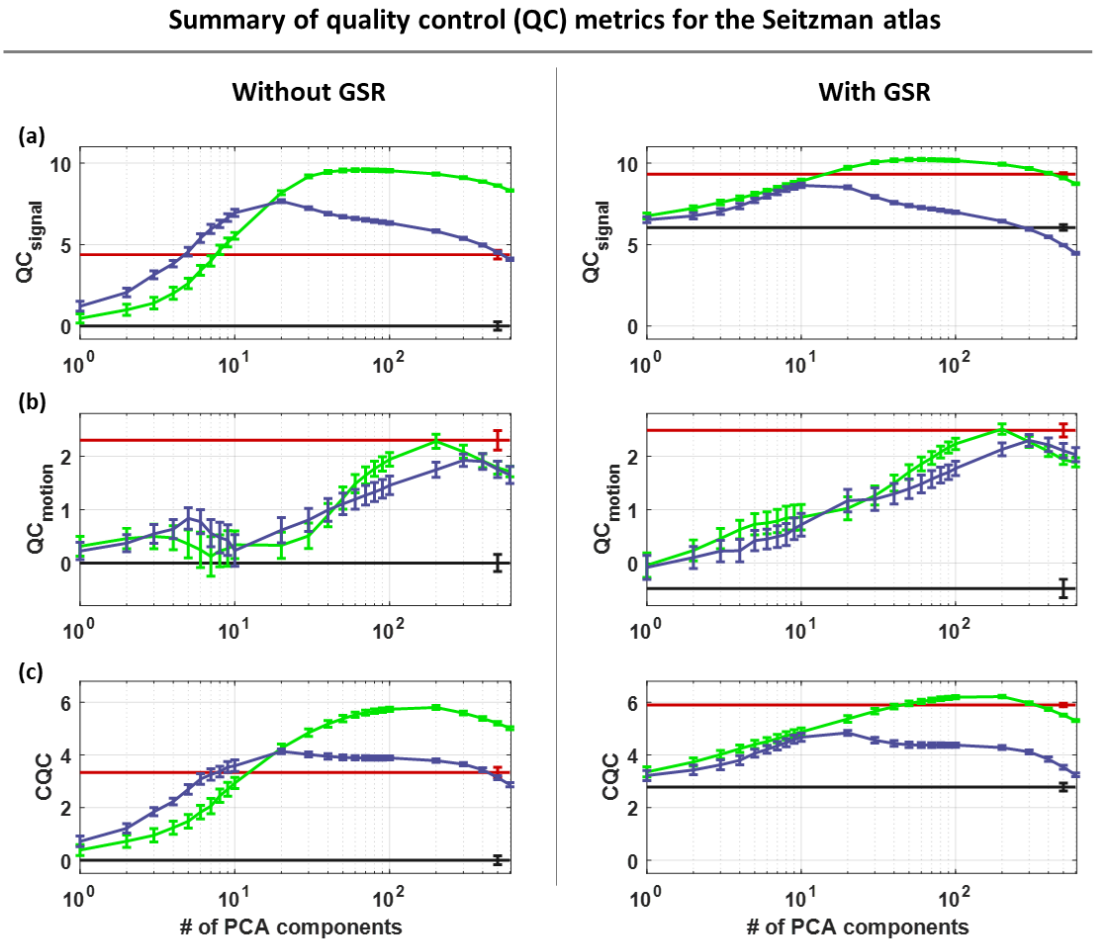
Suppl. Fig. 1. Quality control (QC) scores for the aCompCor analysis, after normalization, using the data in the Gordon parcel space. To summarize all the QC metrics to signal-related and motion-related metrics it was important that the obtained scores from each group of subjects were transformed to z-scores as described in Section 2.8. For both signal- and motion-related metrics, the higher was a normalized score the better the quality of the data was considered. Among all metrics, FCC and FDDVARS demonstrated the most significant improvement in quality with respect to the raw data. For the correspondence of the different curves and lines please see Fig. 2.



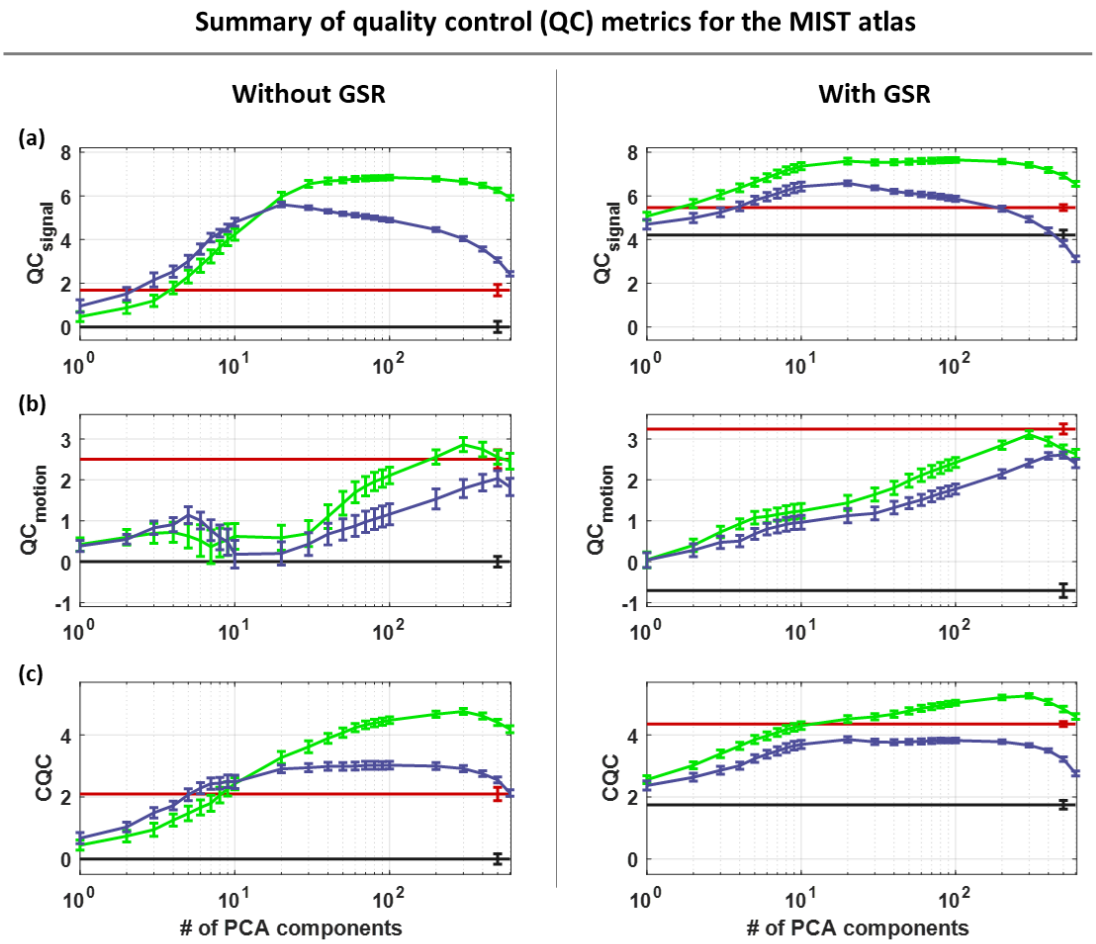
Suppl. Fig. 2. Quality control (QC) scores for the aCompCor analysis using the data in the Seitzman parcel space. In similar to the data at the Gordon parcel space, different trends were observed among the nine QC metrics. Furthermore, the two QC scores FCC and ICCC that are based on the contrast in the FC and ICC matrices, possibly due to the different number of parcels and networks between the atlases, exhibited different range of scores for the data in the Seitzman parcel space compared to the data in the Gordon parcel space. However, the trends of these two scores for varying number of components were similar. For the correspondence of the different curves and lines please see Fig. 2.



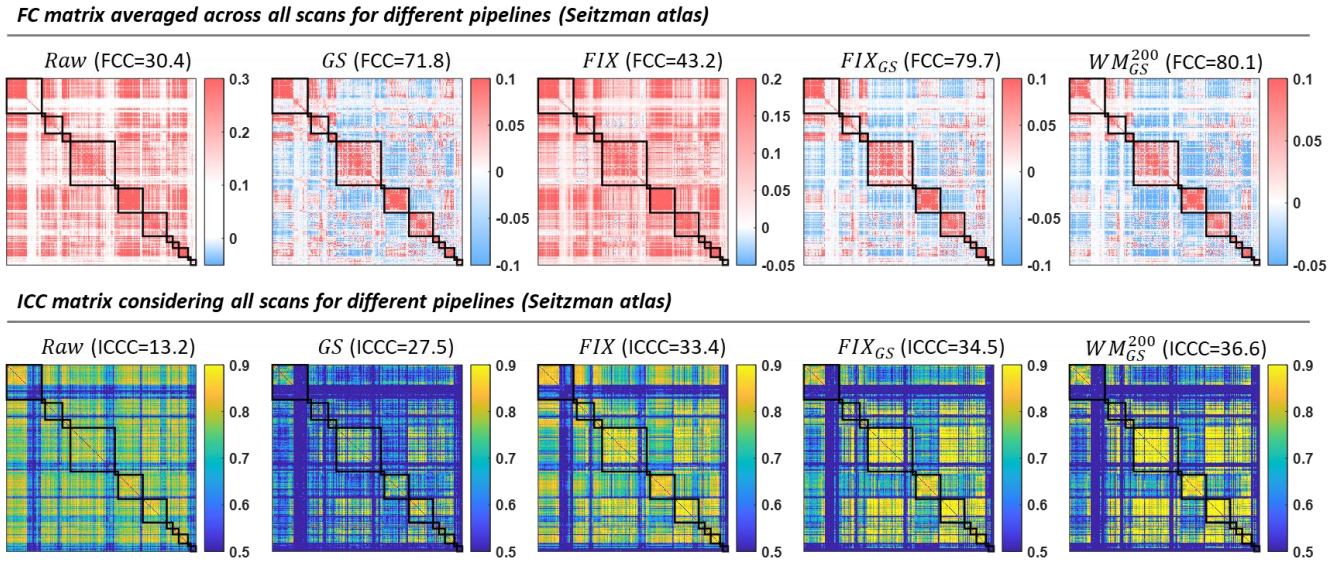
Suppl. Fig. 3. Quality control (QC) scores for the aCompCor analysis using the data in the MIST parcel space. In similar to the data at the Gordon parcel space, different trends were observed among the nine QC metrics. Furthermore, the two QC scores FCC and ICCC that are based on the contrast in the FC and ICC matrices, possibly due to the different number of parcels and networks between the atlases, exhibited different range of scores for the data in the MIST parcel space compared to the data in the Gordon parcel space. However, the trends of these two scores for varying number of components were similar. For the correspondence of the different curves and lines please see Fig. 2.



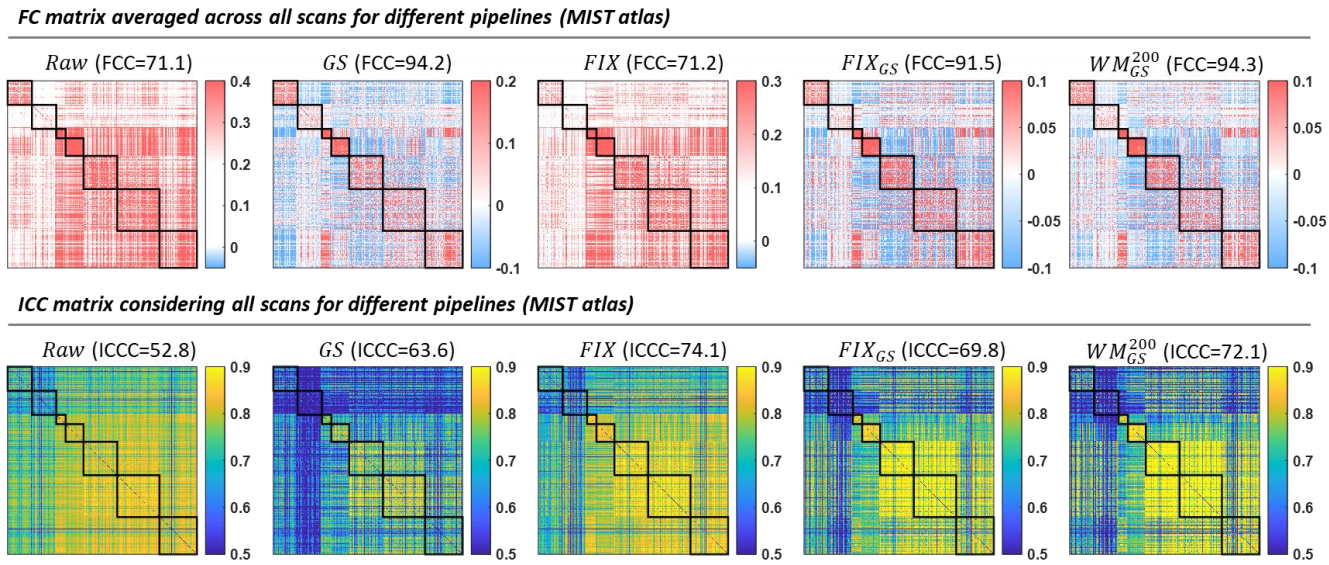
Suppl. Fig. 4. Summarized QC scores for the aCompCor analysis using the data in the Seitzman parcel space. Similar to the data in the Gordon parcel space (Fig. 3), GSR and white matter denoising with 50 to 100 PCA regressors yielded the highest scores for QC_{signal} whereas the more aggressive set of regressors WM_{GS}^{200} achieved the highest score in QC_{motion}. Overall, CQC score that accounts for both QC_{signal} and QC_{motion} was maximized when the set WM_{GS}^{200} was used.



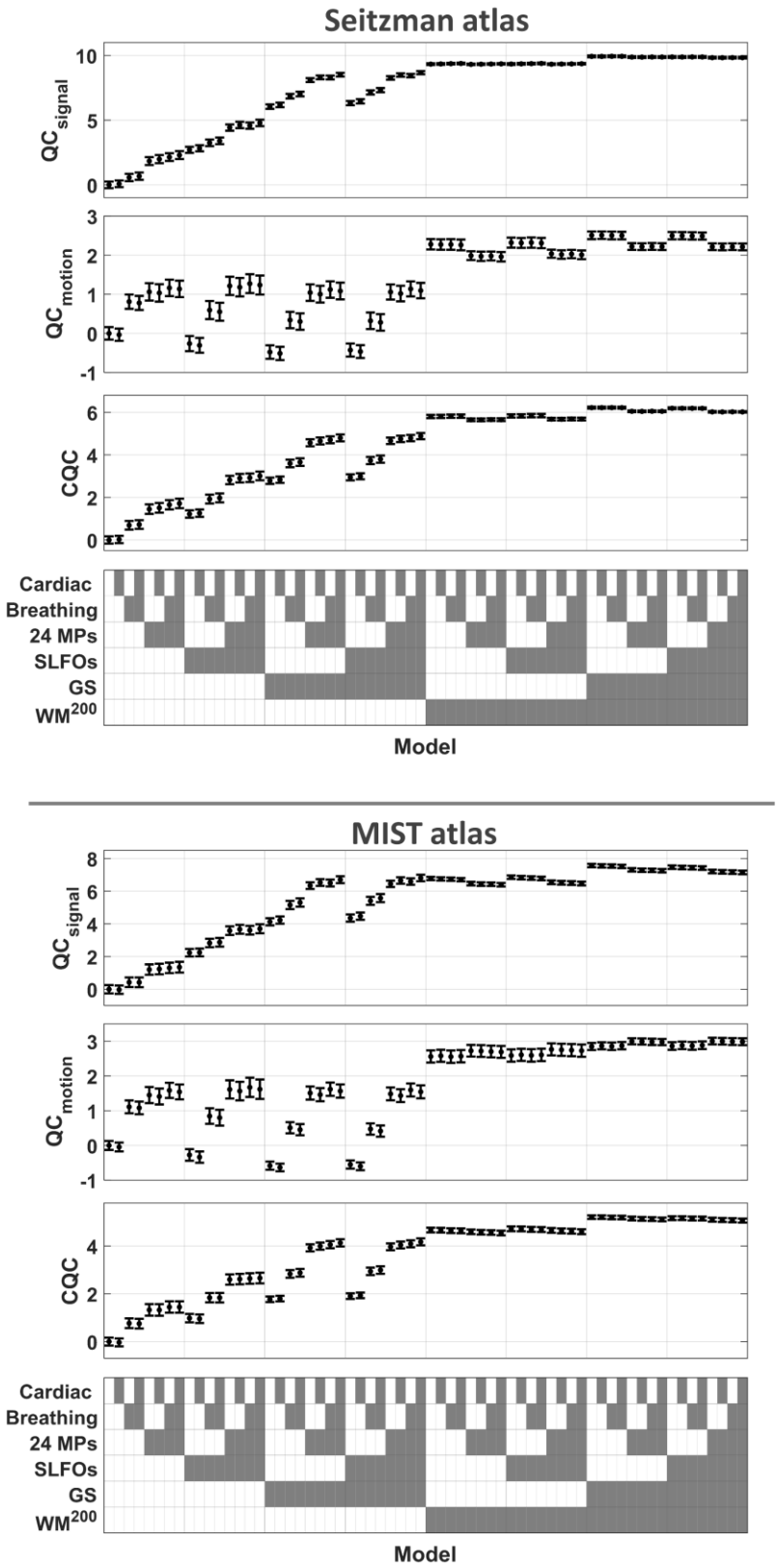
Suppl. Fig. 5. Summarized QC scores for the aCompCor analysis using the data in the MIST parcel space. Compared to the data in the Gordon and Seitzman parcel space, the QC_{signal} was kept relatively stable at a maximum score for a larger range of sets (WM_{GS}^{10} - WM_{GS}^{200}). Moreover, the QC_{motion} yielded a maximum score for the set WM_{GS}^{300}. As a result, the set WM_{GS}^{300} exhibited the highest score for the CQC as well.



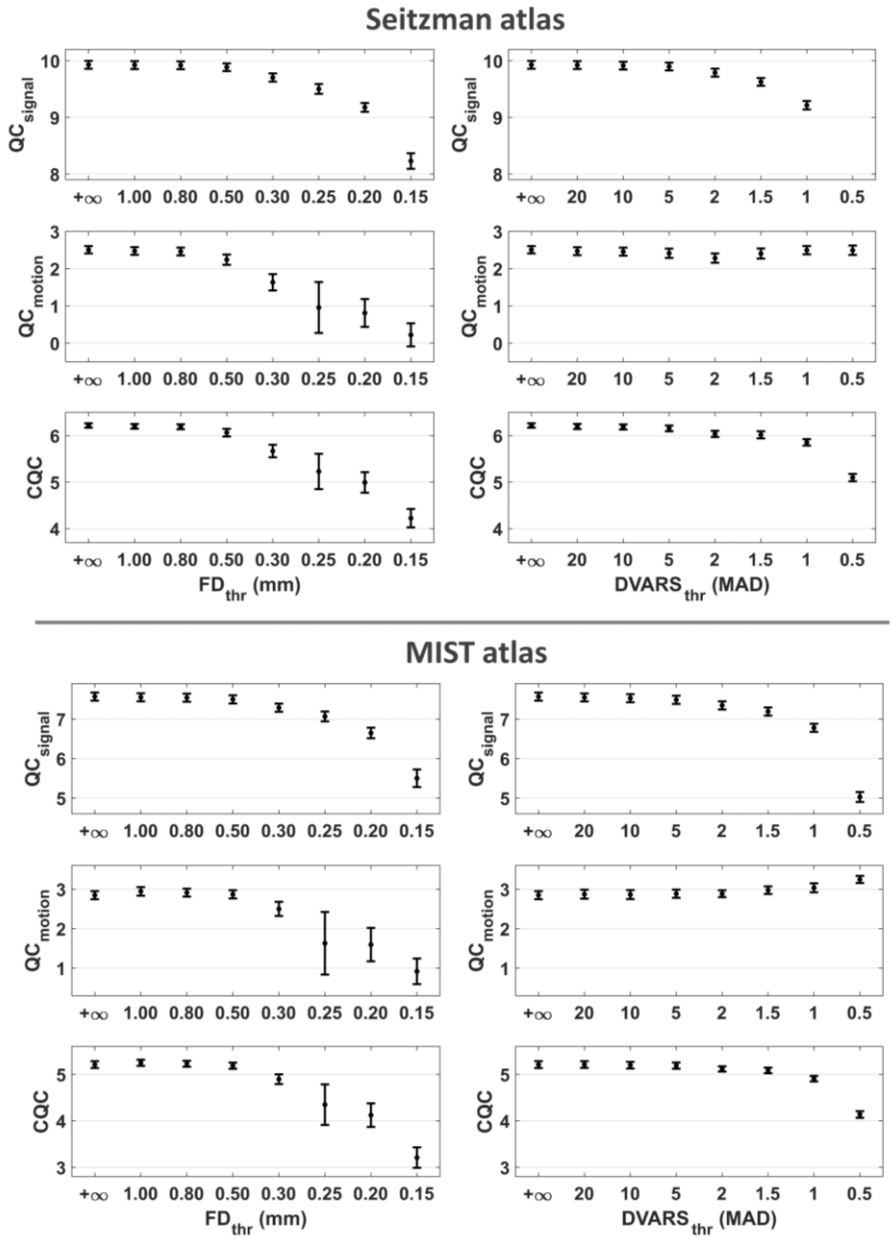
Suppl. Fig. 6. FC (top) and ICC (bottom) matrices considering all scans for different pipelines obtained from the data in the Seitzman parcel space. The pipelines WM_{GS}^{200} and FIX_{GS} significantly improved the identifiability of the networks. Note that many parcels appeared at the end of each network illustrated low correlation and ICC values. These parcels correspond to subcortical parcels and as reported by Seitzman et al. (2018), those parcels demonstrated low temporal signal-to-noise (SNR) in the HCP data which may explain the low correlation and ICC values observed here. Similar to the data in the Gordon parcel space, a large number of BNEs, and especially edges corresponding to interactions between the default mode and fronto-parietal networks, exhibited low FC values but high ICC values.



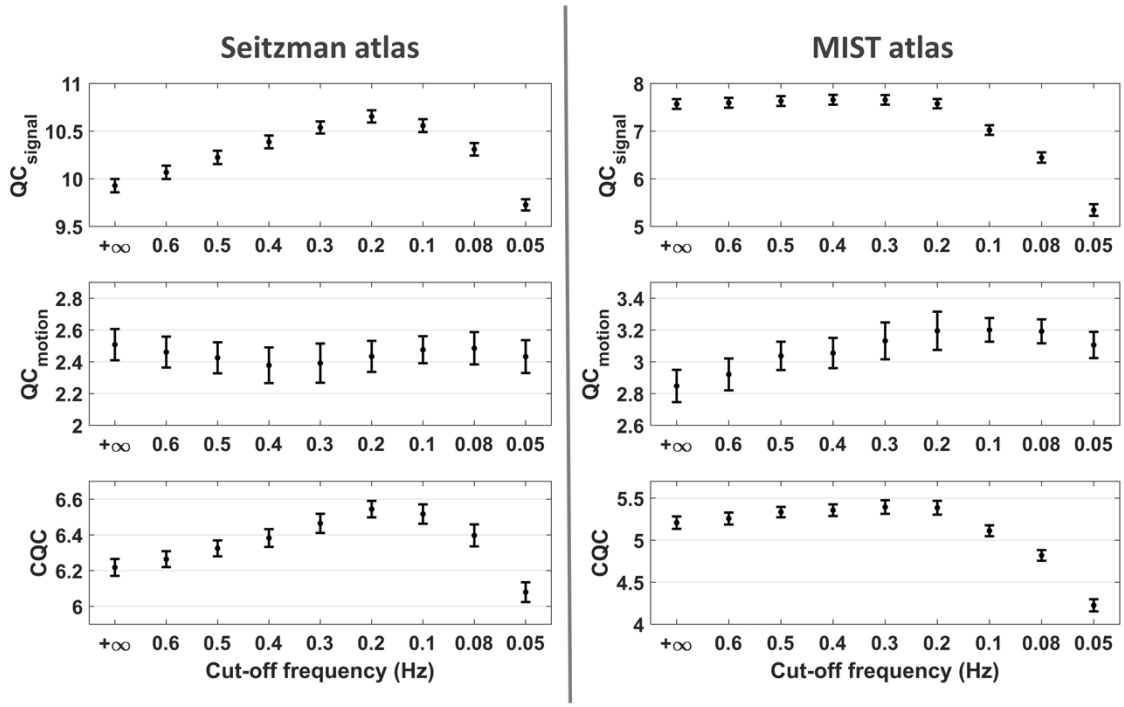
Suppl. Fig. 7. FC (top) and ICC (bottom) matrices considering all scans for different pipelines obtained from the data in the MIST parcel space. Based on the FCC scores obtained from the group-level FC matrices, the pipelines WM_{GS}^{200} and FIX_{GS} significantly improved the identifiability of the networks. However, similar FCC score was observed when only the GS was regressed out. On the other hand, as seen in Suppl. Fig. 2a, at a scan-basis analysis, the FCC scores between the pipelines GS, FIX_{GS} and WM_{GS}^{200} presented significant differences. Moreover, we observe that, sSimilar to the data in the Gordon parcel space, a large number of BNEs, and especially edges corresponding to interactions between the default mode and fronto-parietal networks, exhibited low FC values but high ICC values.



Suppl. Fig. 8. Evaluation of model-based NCTs using the fMRI data in the Seitzman (top) and MIST (bottom) parcel space. As with the data in the Gordon parcel space (Fig. 8), when the model-based regressors related to SLFOs, head motion and breathing motion were used without tissue-based regressors, the data quality as assessed with the three QC metrics QC_{signal}, QC_{motion} and CQC, was improved. However, when the set of nuisance regressors included the GS and the 200 components from WM, none of the model-based regressors was found to provide any additional improvement on the data quality.



Suppl. Fig. 9. Effect of scrubbing in data quality for different threshold values on fMRI data in the Seitzman (top) and MIST (bottom) atlas. When the data were preprocessed with the set of regressors WM_{GS}^{200}, scrubbing before the removal of the regressors did not provide any improvement in the combined summarized QC metric CQC. In contrast, thresholds of FD_{thr} below 0.50 mm led to a significant decrease of the CQC score. In the case of the DVARS, the CQC score was decreased when the threshold was below 1.5 MAD. However, typically, higher values of $DVARS_{thr}$ are used in the literature.



Suppl. Fig. 10. Effect of low-pass filtering in data quality for different cut-off frequencies on fMRI data in the Seitzman (left) and MIST (right) atlas. For both atlases, the highest CQC score was achieved when low-pass filtering was done with a cut-off frequency of 0.2 Hz. The improvement in CQC score was more pronounced in the case of the Seitzman atlas compared to the MIST atlas which was attributed to a substantial incline in the QC_{signal}.