bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

Non-replication of functional connectivity differences in ASD: a multi-site study

Ye He*, Lisa Byrge, Daniel P Kennedy*

Department of Psychological and Brain Sciences, Indiana University Bloomington,
Bloomington, IN 47405, USA

* correspondence can be addressed to either author: he33@iu.edu or dpk@indiana.edu

Abstract

A large and rapidly growing number of studies on autism spectrum disorder (ASD) have used resting-state fMRI to identify various alterations of functional connectivity (FC), with the hope of identifying clinically useful biomarkers or neural mechanisms underlying ASD. However, results have been largely inconsistent across studies, and there is therefore a pressing need to determine the primary factors influencing this lack of replicability. Here, we used resting-state (rs-fMRI) data from the Autism Brain Imaging Data Exchange (ABIDE-I and II) to investigate two factors thought to strongly influence replicability: the use of different denoising methods (i.e., data preprocessing) and data site (which differ in terms of sample, data acquisition parameters, etc.). Using four independently acquired datasets and 31 different denoising strategies, we examined the replicability of patterns of both group-averaged functional connectomes and group-level differences (i.e., ASD vs. control). Within datasets, both group-averages and group-comparisons of functional connectomes were highly consistent ($r = 0.92 \pm 0.06$ / $r = 0.80 \pm 0.09$) across different pipelines, with the largest differences across pipelines reflecting whether global signal regression (GSR) was used. However, across datasets, while group-averages were still highly consistent ($r = 0.86 \pm 0.02$), group differences did not replicate, regardless of denoising strategy: indeed, consistency of group differences was so low ($r = 0.09 \pm 0.05$) that differences identified in one dataset had essentially no relationship with those in other datasets. Across-site similarity remained low even when considering the data at the network (as opposed to edge) level. Because there are a number of additional methodological factors that can influence the reliable detection of group differences (e.g., scanner-related differences, subject differences, post-processing analysis, effect sizes of ASD alterations, amount and quality of data), these results cannot completely rule out the existence of replicable resting-state FC differences in ASD. However, they do highlight the importance of examining replicability in future studies of ASD, and, more generally, call for extra caution when describing and interpreting alterations in functional connectivity across groups of individuals.

Keywords: Autism, resting-state fMRI, functional connectivity, replication

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder with heterogeneous etiology and phenotypic expression. Resting-state functional Magnetic Resonance Imaging (rs-fMRI) -- in which the temporal coupling of spontaneous activity across the brain, or functional connectivity (Biswal et al., 1995; Greicius et al., 2003), is measured -- has been widely used to study differences in functional brain organization in ASD, with hopes of revealing underlying neural mechanisms or identifying FC-based biomarkers (Abraham et al., 2017; Yahata et al., 2016). However, findings of FC alterations in ASD have been highly variable across studies (Hull et al., 2016). This variability of findings may reflect the variability across numerous study-specific factors, including strategies for denoising the data (i.e., preprocessing pipelines) and a host of differences across sites. Yet, without replicable findings that generalize beyond a single study, the utility of rs-fMRI for identifying mechanisms or serving as biomarkers of ASD is uncertain and remains to be demonstrated.

One potential source of variability across rs-fMRI studies has been the methods used for data preprocessing. The blood oxygenation-level dependent (BOLD) signal, while sensitive to changes related to brain activity, is also highly vulnerable to head motion and physiological noise, which can spuriously influence measures of functional connectivity and ultimately affect conclusions from functional connectivity studies (Power et al., 2012; Power et al., 2014; Satterthwaite et al., 2012; Van Dijk et al., 2012; Yan et al., 2013a). Ideally, effective data preprocessing methods would minimize the influence of such nuisance signals and improve reproducibility. Best practices for denoising methods are still evolving and a consensus has yet to be reached, in part because our understanding of how such artifacts influence the BOLD signal remains incomplete (Birn, 2012; Byrge and Kennedy, 2018b; Power et al., 2017). These differences presumably contribute in part to inconsistencies across studies -- different strategies have been used both within and across labs, adding additional uncontrolled and unaccounted for variation in the research literature. Even when researchers attempt to conduct *post hoc* analyses to try to understand how different preprocessing steps could account for study-level differences in ASD, the lack of a ground truth upon which to evaluate measurement accuracy limits our ability to interpret such differences (Müller et al., 2011).

Most common denoising approaches rely on linear regression, whereby various estimates of noise are regressed from the BOLD data. The numerous variations of this strategy come from different choices of which noise estimates to use as regressors. Those most commonly used include measures of head displacement along six translational and rotational dimensions, as well as time series from white matter (WM) and cerebrospinal fluid (CSF). An especially controversial nuisance regressor is the global fMRI signal; proponents of global signal regression (GSR) argue for its efficacy in removing physiological noise (Birn, 2012; Byrge and Kennedy, 2018b; Power et al., 2017), while the concerns include removal of real neural signals (Scholvinck et al., 2010) and distorting clinical group comparison (Gotts et al., 2013; Yang et al., 2014). An additional preprocessing step that can be used in parallel is volume censoring (or "scrubbing"; (Power et al., 2012), in which specific time points associated with excessive amounts of framewise displacement (corresponding to moments of head movement) and/or changes in global signal are excluded from analysis. A related choice is called “spike regression”, which regresses from the data one or more nuisance regressors labeling time points contaminated with excessive motion (Lemieux et al., 2007; Satterthwaite et al., 2013).

Several recent studies have evaluated the performance of different denoising strategies. Although no relationship between motion and functional connectivity should remain following an optimal denoising procedure, these studies found that the strength of residual relationships between FC and artifacts varied widely across commonly-used pipelines (Byrge and Kennedy,

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

2018b; Ciric et al., 2017; Parkes et al., 2018). Given that greater in-scanner head movement is commonly observed in ASD and other clinical populations, differences in preprocessing choices and particularly how those choices deal with artifacts arising from head movement could be a potential source of inconsistent results across rs-fMRI studies. For example, Gotts and colleagues (2013) compared the effects of pipelines with and without GSR on group comparisons of functional connectivity between ASD and controls. They found that group differences varied across pipelines and demonstrated that GSR affected group comparison results. Jones et al. (2010) also found that the use of GSR influenced findings of group differences in connectivity in ASD. Parker and colleagues (2018) systematically evaluated the influence of numerous denoising pipelines on group differences in functional connectivity in schizophrenia. They found that significant group differences were only found in some pipelines (including GSR and aCompCor) and that the overlap between functional connections (i.e., edges) identified in different pipelines was generally low. These findings demonstrate clearly that the choice of denoising pipeline can affect the results of clinical comparisons, including both the presence or absence of group differences and their specific details (e.g., specific edges affected).

Further complicating the picture is that data site effects, or variation across different scanning sites, have been reported in several studies of both task-based and resting-state fMRI (Brown et al., 2011; Dansereau et al., 2017; Noble et al., 2017; Turner et al., 2013; Yamashita et al., 2019; Yan et al., 2013b; Yu et al., 2018). Different data sites present many potential sources of variation, including differences in participant (i.e., cohort) characteristics, image acquisition parameters, scanners, scan procedures, and more. Such uncontrolled variation could undermine the generalizability of results and efforts to uncover underlying mechanisms and clinically useful biomarkers. Clinical and etiological heterogeneity within the ASD population could also exacerbate these difficulties. Nair et al. (2018) compared a local measure of functional connectivity (ReHo, regional homogeneity) between ASD and controls from different samples. They found few consistent results across samples, even when using the same analysis pipeline and examining only data collected with eyes open. They suggested that extra caution should be paid to between-site variability when using multi-site data. However, a recent study reported reproducible ASD-associated alterations of functional connectivity across four large ASD cohorts (Holiga et al., 2019). These two studies both used data from Autism Brain Imaging Data Exchange (ABIDE), a large database aggregating ASD neuroimaging data from multiple sites. Many studies have used ABIDE data by combining data sites to achieve a large sample size rather than examining replication across sites. Therefore, it is necessary to evaluate the replication of results across individual data sites (Brown et al., 2011).

Here, we examined the effects of different denoising pipelines and data sites on the replication of group-level case-control differences in ASD, by using multiple independent datasets within ABIDE. We investigated the consistency across pipelines and across data sites of average functional connectomes within each group, and differences between ASD and control groups, at both region of interest (ROI)-level and large-scale network level.

Methods

Participants

All data analyzed in this study were selected from ABIDE I and ABIDE II (Di Martino et al., 2017; Di Martino et al., 2014). We chose four sites to analyze (NYU, SDSU, UCLA, and UM) because of their large sample sizes and overlapping participant ages (Table 1 and Figure S1). Within each site, participants were included on the basis of mean framewise displacement (FD) no larger than 0.2 mm, manually checked good quality of anatomical images, and age ranging from 10 to 20 years old. To equate groups on FD, control participants at each site were

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license.

selected based on the smallest difference in FD to each ASD participant; no ASD participants were excluded in this procedure. Within each site, there were no group differences in age or mean FD, though some numerically small differences existed between sites (see Table 1). For select additional analyses, we also applied a more liberal motion threshold (mean FD <= 0.3 mm) to include 28% more participants (see Supplementary Table S1 and Figure S1).

Table 1. Demographic information

	NYU (ASD/Control)	SDSU (ASD/Control)	UCLA (ASD/Control)	UM (ASD/Control)
NO.	36/36	34/32	26/26	26/26
Age (Mean ± STD)	13.27 ± 2.61/ 13.93 ± 2.57	14.42 ± 2.52/ 14.31 ± 2.13	14.05 ± 2.19/ 13.58 ± 1.49	14.72 ± 1.77/ 14.92 ± 2.53
Mean FD (Mean ± STD)	0.13 ± 0.04/ 0.13 ± 0.04	0.11 ± 0.04/ 0.10 ± 0.04	0.13 ± 0.04/ 0.12 ± 0.03	0.13 ± 0.03/ 0.13 ± 0.03

Note: Age of the ASD group in NYU was lower than that in UM, and age of the control group in UCLA was lower than that in UM. Mean FD of both the ASD group and the control group in SDSU was lower than their respective groups in NYU and UM (*ps* <= 0.05, uncorrected).

Image Preprocessing

The rs-fMRI scanning parameters for each site are shown in Table 2. All the images were preprocessed using Matlab (R2018a) code made available from a recent study (Parkes et al., 2018) that integrates SPM 12, FSL (FMRIB's Software Library; Smith et al., 2004) and Advanced Normalization Tools (ANTs; Avants et al., 2008)). The T1 images were preprocessed using the following steps: neck removal; segmentation of white matter (WM), cerebral spinal fluid (CSF), and grey matter (GM); five times erosion of WM mask and two times erosion of CSF mask; nonlinear registration of T1 images to MNI space, and applying the transformation to WM, CSF, and GM masks.

Preprocessing of functional images included several steps shared across different denoising pipelines, including the following: removing the first four volumes; slice-timing correction; head motion correction by volume realignment; co-registration to the native structural image using rigid-body registration, and then to the MNI template using nonlinear transformations derived from T1 registration; removing linear trends; normalization of global mean intensity to 1000 units; conducting different denoising strategies (detailed in the next section); bandpass filtering (0.008 - 0.08 Hz); and spatial smoothing with a 6 mm full-width at half-maximum filter.

Table 2. rs-fMRI scanning parameters

	NYU	SDSU	UCLA	UM
Scanner	Siemens 3T Allegra	GE 3T MR750	Siemens 3T TrioTim	GE 3T Signa
TR/TE	2000/15	2000/28	2000/15	2000/30
FA	90	90	90	90
Resolution	3×3×4	3.4×3.4×3.4	3×3×4	3.4×3.4×3

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Volumes	180	180	120	300
Matrix	64×80×33	64×64×42	64×64×34	64×64×40

Denoising Pipelines

We analyzed several commonly-used denoising methods (see details in Parkes et al., 2018), combining different nuisance regression models and volume censoring approaches for a total of 31 denoising pipelines examined (Table 3).

Table 3. Compositions of Denoising Pipelines

Denoising Pipelines	Head motion parameters	Tissue-based Regressors	GSR	Censoring
6H	6			
12H	12			
24H	24			
6H+2W	6	mean WM/CSF		
12H+2W	12	mean WM/CSF		
24H+2W	24	mean WM/CSF		
24H+4W	24	4 mean WM/CSF		
24H+8W	24	8 mean WM/CSF		
6H+aCC	6	aCompCor		
12H+aCC	12	aCompCor		
24H+aCC	24	aCompCor		
6H+2W+Spike	6	mean WM/CSF		Spike
6H+2W+Scrub	6	mean WM/CSF		Scrub
12H+2W+Spike	12	mean WM/CSF		Spike
12H+2W+Scrub	12	mean WM/CSF		Scrub
24H+2W+Spike	24	mean WM/CSF		Spike
24H+2W+Scrub	24	mean WM/CSF		Scrub
6H+2W+GSR	6	mean WM/CSF	1	
12H+2W+GSR	12	mean WM/CSF	1	
24H+2W+GSR	24	mean WM/CSF	1	
24H+4W+GSR	24	4 mean WM/CSF	1	
24H+8W+4GSR	24	8 mean WM/CSF	4	
6H+aCC+GSR	6	aCompCor	1	
12H+aCC+GSR	12	aCompCor	1	
24H+aCC+GSR	24	aCompCor	1	
6H+2W+GSR+Spike	6	mean WM/CSF	1	Spike
6H+2W+GSR+Scrub	6	mean WM/CSF	1	Scrub
12H+2W+GSR+Spike	12	mean WM/CSF	1	Spike
12H+2W+GSR+Scrub	12	mean WM/CSF	1	Scrub
24H+2W+GSR+Spike	24	mean WM/CSF	1	Spike
24H+2W+GSR+Scrub	24	mean WM/CSF	1	Scrub

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Regression of head motion parameters

Head motion parameters are based on six time series reflecting in-scanner head movements along three translational axes and three rotational axes. We examined three variants: 6H (just these original 6 motion parameters), 12H (including the original 6H, plus the first derivative of each as computed by backward differences), and 24H (including 12H, plus the squares of each of the 12 parameters) (Satterthwaite et al., 2013).

Regression of signals from white matter and cerebrospinal fluid

We used two methods to estimate WM and CSF signals: (a) mean WM/CSF, the average time series across voxels within WM and CSF masks, with three variants: mean WM and CSF alone (2W), or adding their temporal derivatives (4W), or adding squares of 4W (8W), and (b) aCompCor, which applies principal component analysis to the time series from WM and CSF voxels separately, and uses the top five principal components for each tissue compartment (Muschelli et al., 2014).

Regression of global mean signal

Global mean signal was calculated by averaging voxel-wise time series across the whole brain (GSR) or extended with squares of it and their temporal derivatives (4GSR).

Volume Censoring

Volume censoring involves censoring specific time points in BOLD data that have excessive head motion, which was evaluated using framewise displacement (FD). We adopted two different censoring strategies: spike regression and scrubbing. To keep consistent with previous work, we calculated FD differently for spike regression and scrubbing and used different thresholds. For spike regression, FD was calculated as the root mean square of framewise changes of six head motion parameters (Jenkinson et al., 2002; Satterthwaite et al., 2013). This FD trace was then used as an additional nuisance regressor in which volumes with FD above 0.25 mm were marked as 1 and otherwise as 0, which was then regressed (together with other regressors) from the BOLD time series. For scrubbing, FD was calculated as the sum of absolute framewise changes of six head motion parameters (Power et al., 2012). Volumes with FD above 0.2mm were excluded from analysis at the end of preprocessing. We excluded subjects with less than 4 minutes of valid BOLD data following spike regression or scrubbing.

Functional Connectome Construction

We used a parcellation template containing 200 cortical ROIs to construct the functional connectome for each subject (Schaefer et al., 2018). Specifically, after preprocessing we weight-averaged the time series of all voxels within each ROI based on their grey matter probability. Then we computed the Pearson's correlation between time series of each pair of 200 ROIs to construct a functional connectivity matrix of each pipeline for each subject, and Fisher-z transformed correlation coefficients for the purpose of normalization.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

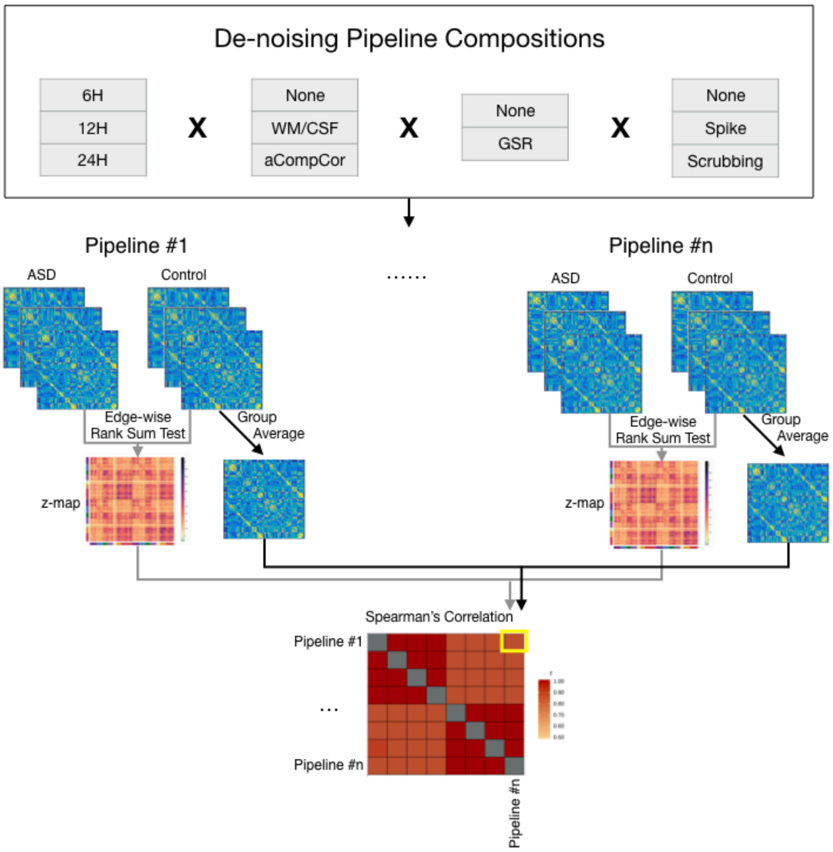


Figure 1. Schematic plot for post-processing analysis within each dataset. We used a total of 31 denoising pipelines, with different combinations of regression of head motion parameters (6H/12H/24H), signals of white matter/cerebral spinal fluid (WM/CSF), global mean signal (GSR), and volume censoring. After preprocessing, functional connectomes were separately constructed with 31 pipelines for each subject. We averaged functional connectomes across each group as well as compared each cell in the connectome between two groups for each pipeline in each dataset. Then we calculated the similarity between group-average functional connectomes, as well as between group-comparison z-maps across pipelines and across datasets.

Assessing Replicability of Whole Functional Connectomes

A schematic is shown in Figure 1 to illustrate our approach. Spearman's correlations were used as the primary similarity metric for assessing replicability. We first averaged functional connectomes across subjects with ASD and across typical controls, separately for each pipeline and for each site. Next, we calculated the Spearman's correlation between group-average functional connectomes for each pair of pipelines to derive a similarity matrix, separately within each data site. To better visualize the distance between pipelines, we used multi-dimensional scaling (MDS) to transform each pipeline-similarity matrix into a representation in two-dimensional space, with each point corresponding to each pipeline and the distance between points corresponding to their degree of dissimilarity. We used Procrustes analysis (without scaling) to best align the plots across sites, using NYU as the reference plot.

We then compared across-site replication within each pipeline, by calculating the Spearman's correlation between each pair of four datasets under each pipeline. We also analyzed root mean squared error (RMSE) as a distance metric (results in Supplementary Material).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Assessing Replicability of Group Differences between ASD and Controls

To evaluate the similarity of group differences across denoising methods and sites, we first compared functional connectivity between each pair of ROIs between the ASD and control groups, using the non-parametric Wilcoxon rank sum tests to reduce the influence of extreme data, after first regressing out age as a covariate. A 200 × 200 z-value matrix was obtained for each pipeline in each dataset. We then analyzed the overall similarity of these z-value maps (z-maps) across pipelines within each dataset, and within each pipeline across datasets, using Spearman correlation and RMSE as described previously. The results described subsequently were similar when using parametric t-tests.

In addition to comparing these z-maps, we also examined only those functional connections (edges) with significant differences between the ASD and control groups. To do this, we used Dice Index (Dice, 1945) to assess the ratio of overlapping edges with significant group differences across those two pipelines or sites, after first binarizing each z-map (i.e. setting to 0 all edges except those with corresponding p-values <= 0.005). Permutation tests were used to examine whether these overlapping ratios were above chance. Specifically, to generate a null distribution while limiting computational demands, we chose only pipelines with 24H to compare. First, we shuffled the diagnostic labels of all the subjects to either ASD or controls for each site, keeping original sample sizes for each group. Then we compared these two new groups to derive a null z-map for each 24H-based pipeline within each site. This procedure was repeated 100 times for each pipeline. Next, we randomly chose one out of the 100 null z-maps from each site and calculated the Dice index between each pair of sites. This procedure was repeated 1,000 times to generate a null distribution of Dice indices for each pipeline. The p-value was decided by the location of the actual Dice index in the null distribution and corrected by the false discovery rate (FDR; Benjamini and Hochberg, 1995).

We also examined the similarity of group differences between data sites at a large-scale network level. We mapped the whole functional connectome to a 7 functional networks template and obtained a 7 × 7 connectivity matrix by averaging connectivity of edges in each cell (Yeo et al., 2011). Using the same statistical method to compare each cell between the ASD and control groups, we obtained a z-map for each pipeline in each site and evaluated similarity across pipelines and sites as described above.

Data and code availability

All data is available from the ABIDE repository, preprocessing code was available from Parkes et al. (2018), and our code is available upon request. This complies with our funding agency requirements.

Results

Group averages and group differences replicate across pipelines within each site

We first compared the similarity of group-average functional connectomes of typical controls across different denoising pipelines separately within each data site. Generally, functional connectomes were highly similar across pipelines within each dataset (Figure 2, top; NYU, $r = 0.92 \pm 0.07$; SDSU, $r = 0.93 \pm 0.06$; UCLA, $r = 0.93 \pm 0.05$; UM, $r = 0.91 \pm 0.07$). Results were similar for group-average functional connectomes in ASD (NYU, $r = 0.93 \pm 0.06$; SDSU, $r = 0.93 \pm 0.06$; UCLA, $r = 0.92 \pm 0.06$; UM, $r = 0.91 \pm 0.07$; Supplementary Figure S1). As is apparent from the quadrant structure of Figure 2, GSR was a major influence on similarity of average functional connectomes across pipelines, such that similarity was extremely high with the same GSR status but reduced when pipelines differed in their use of GSR. We used multi-

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

dimensional scaling to represent this graphically (Figure 2, bottom), which demonstrates that the use of GSR is a primary dimension upon which results are either similar or different from one another. RMSE results were consistent with the correlation results (Supplementary Figure S2).

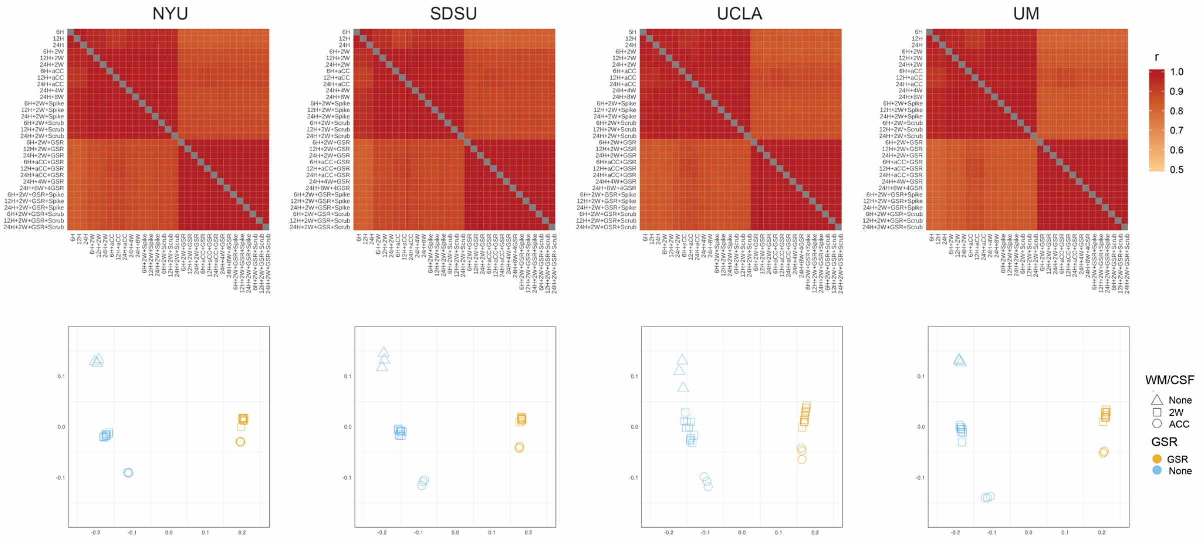


Figure 2. Consistency of group-average functional connectome across pipelines. The upper panel shows Spearman's correlation coefficient of average functional connectomes across pipelines. It indicates high similarity across pipelines, though pipelines with different GSR status were less similar, as is seen in quadrant structure. The bottom plots provide a different visualization of relative similarity among different pipelines based on multi-dimensional scaling. Each data point represents a pipeline (note that not all points are visible because there is a high degree of overlap between some of them). It directly shows the major factor differentiating pipelines is based on the usage of GSR. The triangle shape corresponds to the basic pipelines (which only regress out 6H/12H/24H), the square shape corresponds to the pipelines adding WM/CSF regression, and the circle shape corresponds to pipelines using aCompCor.

Then, we analyzed similarity between pipelines of ASD-Control comparisons of functional connectomes, within each site. The results were also consistent across pipelines within each dataset (NYU, $r = 0.81 \pm 0.09$; SDSU, $r = 0.81 \pm 0.09$; UCLA, $r = 0.80 \pm 0.09$; UM, $r = 0.79 \pm 0.11$). As in Figure 2, Figure 3 shows that GSR was also a dominant factor in similarity of group differences across pipelines – highly similar results with concordant use or non-use of GSR (i.e., either both present or absent), but reduced similarity when pipelines were discordant in their use of GSR (concordant: $r = 0.87 \pm 0.06$; discordant: $r = 0.72 \pm 0.04$, averaged across four sites). The Dice index, measuring the ratio of overlapping significant edges ($p <= 0.005$, uncorrected) between pipelines, also supported the correlation results, with high values between pipelines with the same GSR status, and low values between pipelines with different GSR status (see Figure 3, bottom row).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

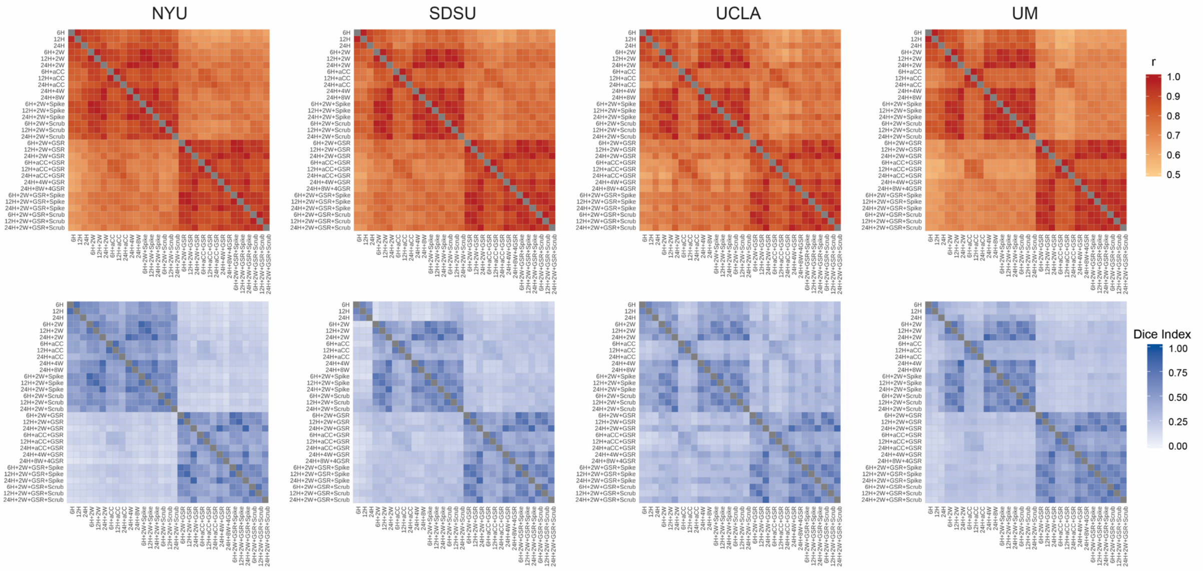
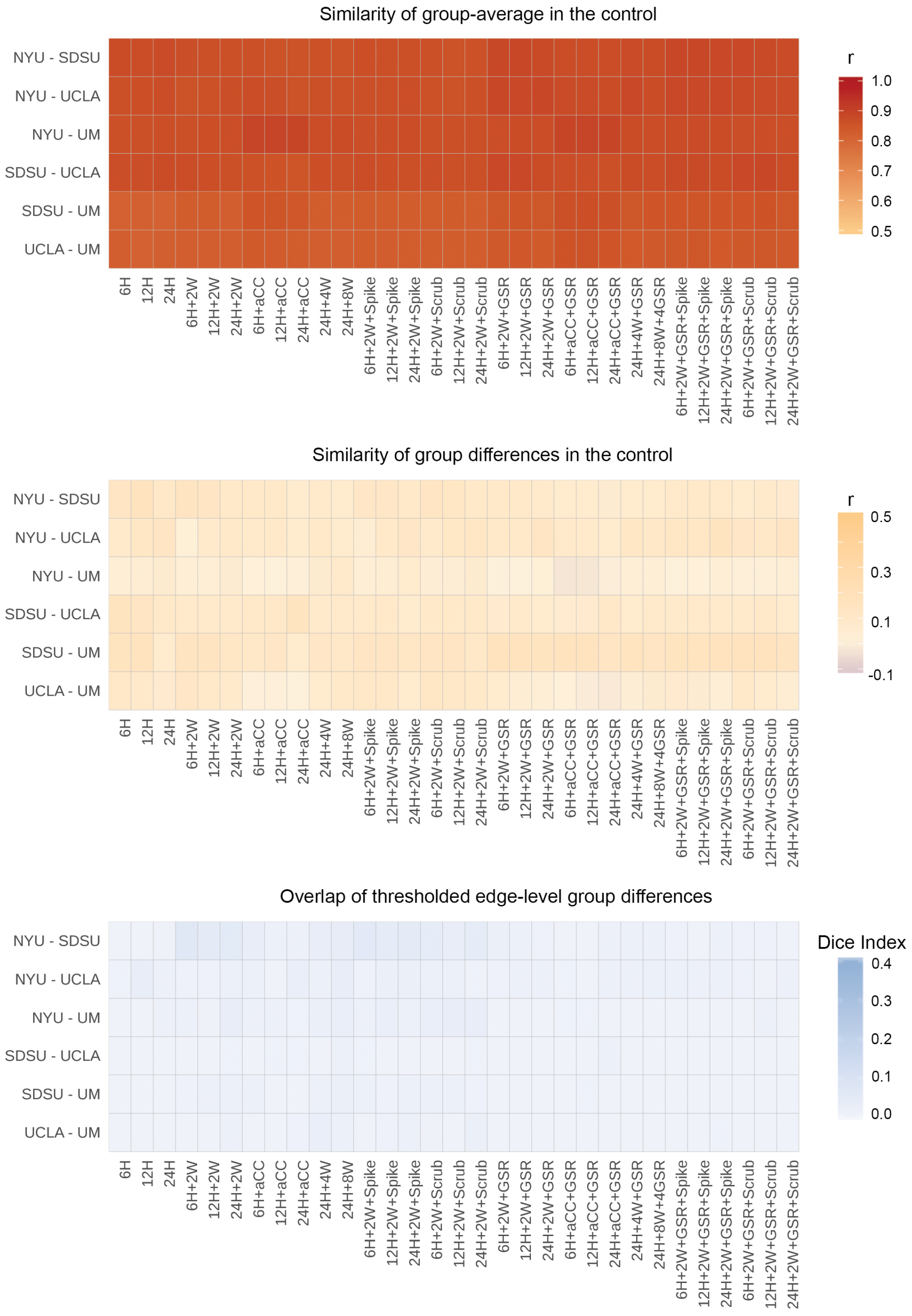


Figure 3. Consistency of group differences in functional connectivity across pipelines, within each data site. The top row shows correlations between group-comparison z-maps for each pair of pipelines. The bottom row shows the Dice index measuring the overlapping across pipelines of edges with significant differences ($p <= 0.005$) between the ASD and control groups. These figures indicate the group-difference patterns were consistent across pipelines, especially within pipelines with GSR or without GSR.

Group averages, but not group differences, replicate across sites

Next, we examined the effect of data site on replicability. We first analyzed the similarity of group-average connectomes of typical controls across data sites, within each pipeline. Figure 4 shows that group-average connectomes are similar across data sites within each pipeline, for all pipelines ($r = 0.86 \pm 0.02$), though data sites were slightly more similar to each other for pipelines with GSR ($r = 0.87 \pm 0.02$) than those without GSR ($r = 0.85 \pm 0.02$; rank-sum test, all *ps* < 0.05, except NYU-UM, corrected). Similar results were obtained with group-average connectomes for the ASD group and using RMSE (Supplementary Figure S3).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



Similarity of group-average in the control

Similarity of group differences in the control

Overlap of thresholded edge-level group differences

Figure 4. Consistency of group-averages and inconsistency of group-comparisons of functional connectomes across data sites. Between each pair of data sites and for each pipeline, there were high correlations of group-average functional connectomes (top) but low correlations of group-comparison z maps (middle). The low Dice indices in the bottom plot indicate that significantly different edges rarely overlap across data sites.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

However, the pattern of group differences in functional connectomes between the ASD and control groups was poorly replicated across data sites (Figure 4 and Figure 5; r = 0.09±0.05), in contrast to the relatively high similarity of group differences across pipelines within the same data sites (cf. Figure 3). Although this low similarity between z-maps for different data sites was observed regardless of pipeline, the effects of GSR on this similarity were mixed depending on specific pairwise site comparisons – correlations were significantly higher for pipelines with GSR compared to without GSR in NYU-UCLA ($z = 2.64$, $p = 0.05$) and SDSU-UM ($z = 4.07$, $p <= 0.001$), and lower in NYU-SDSU ($z = -3.06$, $p = 0.01$), NYU-UM ($z = -3.38$, $p = 0.004$), SDSU-UCLA ($z = -2.78$, $p = 0.03$), and UCLA-UM ($z = -2.55$, $p = 0.06$) (each was corrected for multiple comparisons). In other words, the effect of GSR was not consistent in either increasing or decreasing across-site similarity of group differences. We further focused only on the specific edges showing significant group differences. This analysis, however, revealed the same pattern of results: Dice indices were still very low across data sites for all pipelines (Figure 4; Dice = 0.01±0.01; range = 0.00 - 0.07). Permutation tests indicated that the Dice indices were not significantly higher than chance when correcting for multiple comparisons. Figure 5 shows the overlap of significant edges across the four sites from several representative pipelines. No edges ever significantly differed in more than two out of the four sites, in any pipeline.

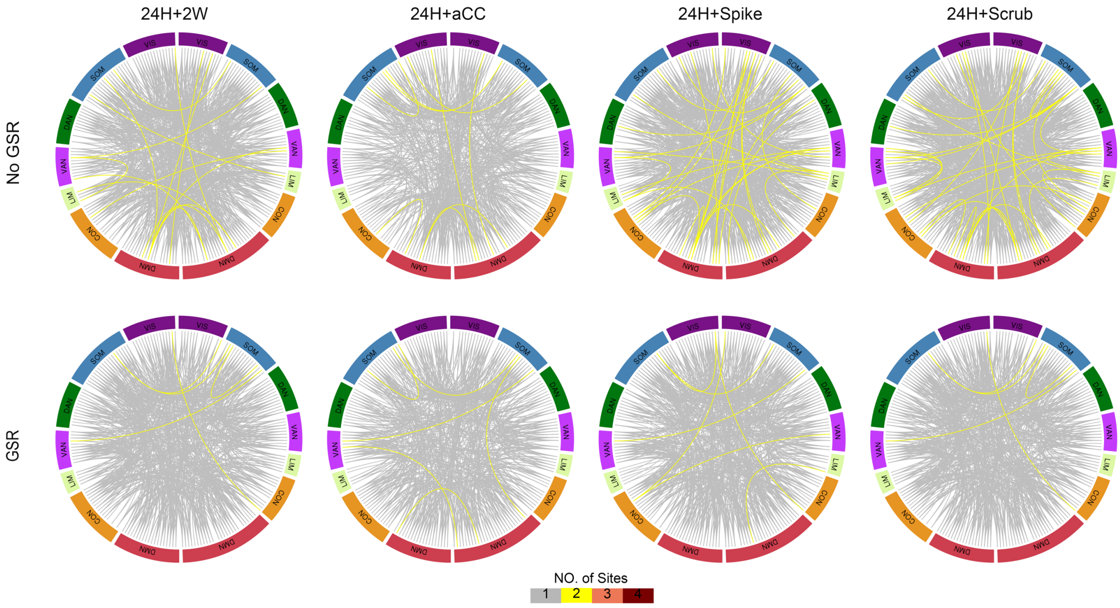


Figure 5. Edges showing significant differences were replicated in at most two of four sites. The circular plots show how many times the edges had significant differences across four sites separately in representative pipelines (24H+2W, 24H+ACC, 24H+2W+Spike, 24H+2W+Scrub, 24H+2W+GSR, 24H+ACC+GSR, 24H+2W+GSR+Spike, and 24P+2P+GSR+Scrub). The grey line indicates that connectivity between two regions was significantly different in one site, while the yellow line indicates the connectivity was significantly different in two sites. No edge appeared more than two times (i.e. more than 50% of sites). VIS: visual network; SOM: somatomotor network; DAN: dorsal attention network; VAN: ventral attention network; LIM: limbic network; CON: control network; DMN: default mode network.

In addition to examining consistency of group differences across data sites at the fine ROI edge-level resolution, we also examined the data at a larger-scale network level.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Correlations between z-maps reflecting ASD-control differences varied across data sites. Figure 6 shows that z-map correlations at the network level (*r* = 0.14±0.24), with none being significant after multiple comparison correction. The effects of GSR on between-site similarity were still mixed depending on the specific pairwise site comparisons -- correlations were higher for pipelines with GSR for NYU-UCLA (*z* = 3.88, *p* <= 0.001), and lower for NYU-UM (*z* = -4.88, *p* <= 0.001) (see Figure 6).

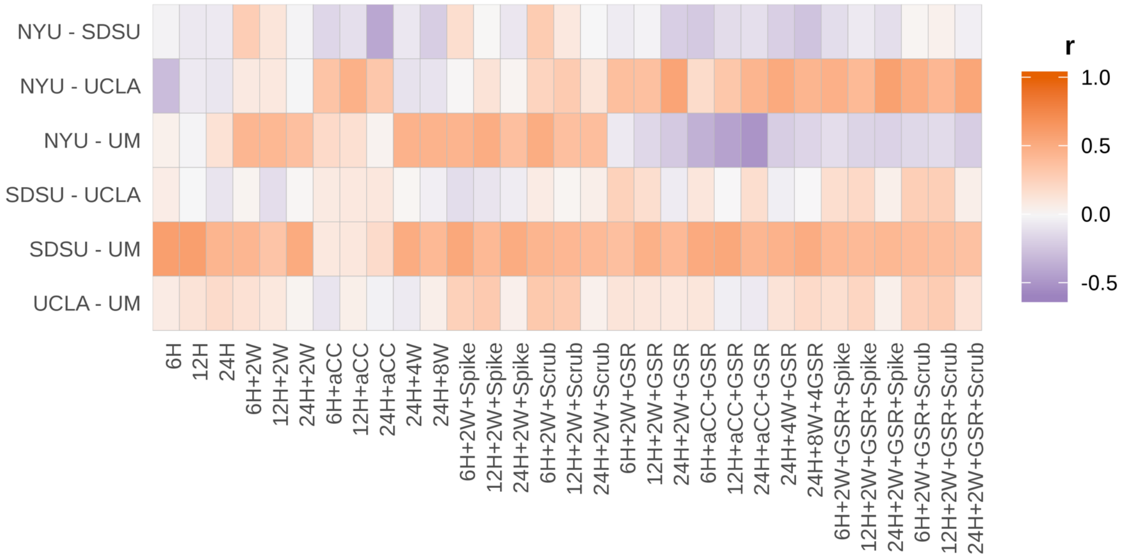


Figure 6. Inconsistency of group differences at the network level across data sites. Note that here, compared to Figure 4 (middle row), the results are more variable.

Discussion

Our findings demonstrate a remarkable lack of replication of group-level differences in resting-state functional connectivity in ASD. This result is largely consistent with the varied and often conflicting published literature in ASD when taken as a whole – for example, even the basic directionality of effects is still debated (i.e., systematic overconnectivity, underconnectivity, both, or neither). However, the source of this inconsistency has remained unknown, because too many factors differ across studies and methodology is free to vary widely with no consensus on best practices. Here, we show that lack of replication cannot be attributed to differences in preprocessing procedures – when we use the same denoising strategy across the four different datasets, we still observed a total lack of replication. Furthermore, this was true regardless of which of the 31 different denoising methods we used – i.e., the degree of replication did not improve in any meaningful way with any particular approach (e.g., GSR vs. not). Importantly, this lack of replication was specific to group-level differences and did not extend to basic connectome architecture – when comparing average connectomes across sites, we found a high degree of similarity, regardless of denoising procedure. Based on these results, we conclude that while preprocessing may still contribute in part to the lack of replication seen across studies (as it certainly adds variability), these differences are likely not the major factor accounting for such inconsistencies and suggest that other factors play a more significant role.

If differences in denoising strategies cannot adequately explain the lack of across-site replication, an important question is what other factors may account for it. There are at least four possibilities: 1) specific scanner/acquisition/procedural differences; 2) subject-level (cohort) differences; 3) differences in post-processing analysis – e.g., the scale or level (region-

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

of-interest, whole connectome, or network levels); 4) small, hard to detect, or even absent differences in functional connectivity in ASD. We unpack these possibilities in the following paragraphs, with each having specific implications for design and analysis of future studies.

On the data collection side, it is possible that factors (including some that remain uncontrolled in the present study) contribute to this lack of replication. These factors include things like scanner and acquisition parameter differences (e.g., pulse sequence, voxel size, phase encoding directions, scanner manufacturer, etc.; Yamashita et al., 2019) as well as experimental procedural differences (e.g., eyes open or closed, immediately experiences preceding the functional scan; Nair et al., 2018). Importantly, if such scanner, acquisition, and procedural factors so strongly influence the ability to detect reliable differences in ASD, then uncoordinated efforts to uncover functional connectivity differences are largely futile. In other words, if results from one site with their particular procedures and parameters are unrelated to the results from another site using their own procedures and parameters, then combining datasets like the efforts undertaken in ABIDE (Di Martino et al., 2017; Di Martino et al., 2014) have limited utility. Indeed, the ability to easily share data across sites was one of the factors underlying initial excitement about resting-state MRI approaches – the "task" was instantly standardized. Fortunately, these factors, while they do contribute to across-site variance, tend to be small in terms of effect size (Brown et al., 2011; Dansereau et al., 2017; Noble et al., 2017) or result in localized differences (Nair et al., 2018), consistent with our finding that group-average connectomes were highly reliable across sites. However, to further increase chances of replication, either *a priori* coordination and standardization of procedures (Glover et al., 2012) or the implementation of post-processing methods designed to increase multisite data harmonization would both be possibilities (Yamashita et al., 2019; Yu et al., 2018).

Another factor related to data collection that potentially underlies our inability to replicate across sites could be subject-level (i.e., cohort) differences or biases (Yamashita et al., 2019). A non-exhaustive list of these factors includes ASD severity, cognitive level, co-morbidities, treatment history and current treatment status (e.g., medication), basic demographic factors including age, sex, race, ethnicity, education, socioeconomic status, and so on. These cohort differences emerge both from practical constraints (e.g., regional biases in terms of participant demographics in different locations) and from the various choices made regarding the recruitment process (e.g., the types of recruitment channels such as clinics vs. communities, and any specific inclusionary and exclusionary criteria). There are several options to remedy these issues. Of course, one could apply tightly specified and standardized criteria to match participants across a host of these factors, but in doing so the generalizability of the findings to the broader ASD condition is reduced. A more practical consideration is that attempting to better match sites on some of these factors can result in smaller sample sizes -- for example, in our study, we excluded 184 participants (nearly 31%) in order to better closely match sites on just one of these factors (age). However, it is not necessarily the case that applying more restrictive criteria is always better than including more participants (Abraham et al., 2017). Another way to proceed is to identify the critical factors or grouping of factors that explain significant variance in the data (Smith et al., 2015), and statistically control for those. Other proposals have suggested increasing sampling diversity by collecting relatively smaller numbers of participants at many different sites, rather than many participants at one site (Dansereau et al., 2017; Yamashita et al., 2019) – indeed, one recent study (Holiga et al., 2019) that reported replicable findings using the ABIDE dataset combined data across multiple sites as opposed to treating each ABIDE site separately as in the present work. Regardless of the approach one uses, accounting for these subject-level differences is likely an important consideration, as recent work has highlighted that subject-level factors explain more variance than site-level factors (Brown et al., 2011; Dansereau et al., 2017; Gountouna et al., 2010; Noble et al., 2017).

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

On the analysis side, it is important to note that our findings of a lack of replication are specific to our particular analyses using both whole connectome ROI-level and a large-scale network level organization, and do not rule out the possible existence of any other replicable group-level effects in ASD. It is very possible that replicable results could be found when considering the very same data at a different scale or resolution, or with that data analyzed in a different way. For example, King and colleagues (King et al., 2018) found replicable atypical temporal dynamics in rs-fMRI timecourses. Holiga et al. (2019) recently found replicable results regarding functional connectivity in ASD across four very large datasets that also included ABIDE data. Other studies have used machine learning approaches to generalize to independently acquired datasets (e.g., Abraham et al., 2017; Yahata et al., 2016). In one of these (Abraham et al., 2017), prediction accuracy was affected by parcellation method, suggesting that replicability may be sensitive to these sorts of analysis choices (e.g., spatial normalization, parcellation). Additionally, different scales of connectivity analysis exhibit different sensitivities and vulnerabilities to site effects (Noble et al., 2017), demonstrating a complex and intertwined relationship between many of the factors discussed above. We should mention, however, that although there are different ways of dividing and grouping the data, these approaches mostly still fundamentally rest on the ability to accurately and reliably measure edge-level differences in ASD (e.g., Yahata et al., 2016). For example, more complex statistical constructs that can be used to compare brain organization between groups (e.g., graph theoretic network measures; He et al., 2018; Rubinov and Sporns, 2010) fundamentally must build upon reliable and replicable measurement of connectomes. Thus, lack of replication as described in the present work should be of concern to researchers.

The final possibility that ought to be considered is that functional connectivity differences in ASD are very small, hard or impossible to detect with current technology, or even non-existent. While hundreds of published studies to date have reported on functional connectivity differences in ASD, the overall lack of consensus is concerning. Surely there are neural differences in brain organization and functioning in ASD, given that it is a neurodevelopmental disorder, but whether or not these differences can be reliably detected using current neuroimaging methodologies remains an open question, especially given the present results. The growing number of studies that now examine and in some cases demonstrate out-of-sample replication provide hope that such signals do in fact exist (Holiga et al., 2019; Yahata et al., 2016). But, because of the above factors and in addition to a host of others (e.g., motion), small differences may be easily obscured (Tyszka et al., 2014).

What does this all mean? The pessimistic view would be that researchers should give up on searching for common group-level effects in ASD. However, we believe that this conclusion would be very premature for a number of reasons. (1) It is possible that effects are heterogeneous across participants, so group-level analysis starting with the assumption of homogeneous groups may be both largely underpowered and not able to fully account for the group level variance. (2) It is possible that improvements in detecting signal in the face of the large amounts of measurement noise that plague resting-state analyses will eventually unmask important group-level differences. In this case, if it is a detection problem, continued advances in acquisition and analysis methodology may get us closer to detecting reliable differences in ASD. (3) Additional experimental procedures can be employed to ensure more reliable estimates of an individual's connectome. For example, collecting more data from each individual participant can reduce measurement noise and ensure greater confidence in the results via within-sample replication (Anderson et al., 2019; Byrge and Kennedy, 2018a; Finn et al., 2015; Nee, 2019), prior to attempting across-site replication.

While our results suggest that lack of replication cannot be solely attributed to differences in denoising procedures (since using the same preprocessing procedures did not increase across-site replication), this does not mean that they are entirely inconsequential. Here,

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

we show that, while there are essentially an unconstrained number of choices for preprocessing, some of these choices have a more significant impact on the results than others (though not necessarily in a consistent way). Figure 2 demonstrates that one of the most significant factors is whether or not GSR is included as a preprocessing step. Its inclusion resulted in slightly more similar group-averaged connectomes across sites -- however, whether more similar group-averaged connectomes is a good thing or not remains unclear. The positive interpretation of this finding is that GSR helps to eliminate measurement noise (Byrge and Kennedy, 2018b; Ciric et al., 2017; Parkes et al., 2018; Power et al., 2014; Power et al., 2017), resulting in more similar connectomes, whereas the less positive interpretation is that GSR eliminates individual variation that might be of interest or distorts group-level differences (Gotts et al., 2013; Scholvinck et al., 2010; Uddin, 2017; Yang et al., 2014). Our results cannot disambiguate these possibilities from one another. Furthermore, in terms of group differences, we found that the effects of GSR on across-site replicability were not consistent, and instead depended on which specific sites were compared to one another (see Figure 4, middle panel, and Figure 6). For some site comparisons, use of GSR significantly increased similarity between them, whereas for others it decreased it, and yet others where it was unchanged, suggesting a complex interaction between the use of GSR and site-level factors.

In addition to the possible factors already discussed above that may limit the detection of reliable group effects, some additional limitations of this study are worth mentioning. One criticism is that correlations between whole connectome z-maps comparing groups is perhaps a relatively insensitive way to examine this data. For instance, a localized difference in a small number of edges or nodes would easily be obscured in the present whole-brain analyses. However, we did also examine only the most significant edges that differed between groups, and also examined data aggregated at the network level – both yielded equally disappointing results. Another limitation is that although we examined 31 different denoising pipelines, these did not include ICA-based methods (e.g., ICA-AROMA, Pruim et al., 2015; FIX, Salimi-Khorshidi et al., 2014); whether these perform any better in terms of across-site replication should be examined in the future. Regardless, while these results do not speak to all possible denoising strategies, they do cover a large swath of variations in methodologies found in the functional connectivity literature (Varikuti et al., 2017). Another limitation of the present study is the relatively small sample sizes. This was a consequence of both matching groups by age and also applying strict quality control (i.e., movement thresholds, anatomical image quality requirements). However, we note that our sample size was sufficiently powered to detect medium-large to large effects within each dataset, suggesting that any differences are likely smaller than this. We also included more subjects (Table S1) by lowering the threshold on head motion (mean FD <= 0.3 mm) to repeat the analysis, and obtained similar results (see Supplementary Figure S4). Finally, the present study included a relatively large age range from 10-20 years, corresponding to a broad neurodevelopmental period spanning childhood through adolescence and into young adulthood. It is possible that more consistent effects would be identified if the age was constrained even further – however, further restricting the range would have reduced the number of sites and subjects that we could have included.

In sum, the present study demonstrated that one’s choice of denoising pipeline is not the main factor underlying the lack of replication of differences in ASD. Instead, the most parsimonious explanation for the lack of replication is that group-level differences are either small or non-existent, and/or swamped by site and sample effects. However, we remain optimistic that continued developments toward improving methodology and approaches will help to eventually reveal reliable patterns of functional connectivity alterations in ASD. These results highlight the need to continue examining reliability of findings going forward, and demonstrate that approaches that improve sensitivity to detect disorder-related alterations are still needed.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Declarations of interest: none

Acknowledgement

This work was supported by the NIH (R01MH110630 and R00MH094409 to DPK), and NICHD (T32HD007475 Postdoctoral Traineeship to LB). The authors acknowledge the Indiana University Pervasive Technology Institute for providing HPC (Big Red II, Karst, Carbonate), visualization, and storage resources that have contributed to the research results reported within this paper. This study was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU was also supported in part by Lilly Endowment, Inc. This material is based upon work supported by the National Science Foundation under Grant No. CNS-0521433.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147, 736-745.

Anderson, A.N., King, J.B., Anderson, J.S., 2019. Neuroimaging in Psychiatry and Neurodevelopment: why the emperor has no clothes. The British Journal of Radiology, 20180910.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med Image Anal 12, 26-41.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Statistical Methodology 57, 289-300.

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

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. Magnetic Resonance in Medicine 34, 537-541.

Brown, G.G., Mathalon, D.H., Stern, H., Ford, J., Mueller, B., Greve, D.N., McCarthy, G., Voyvodic, J., Glover, G., Diaz, M., Yetter, E., Ozyurt, I.B., Jorgensen, K.W., Wible, C.G., Turner, J.A., Thompson, W.K., Potkin, S.G., Function Biomedical Informatics Research, N., 2011. Multisite reliability of cognitive BOLD data. Neuroimage 54, 2163-2175.

Byrge, L., Kennedy, D.P., 2018a. High-accuracy individual identification using a “thin slice” of the functional connectome. Network Neuroscience 3, 363-383.

Byrge, L., Kennedy, D.P., 2018b. Identifying and characterizing systematic temporally-lagged BOLD artifacts. Neuroimage 171, 376-392.

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.

Dansereau, C., Benhajali, Y., Risterucci, C., Pich, E.M., Orban, P., Arnold, D., Bellec, P., 2017. Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. Neuroimage 149, 220-232.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J.S., Assaf, M., Balsters, J.H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L.M.E., Bookheimer, S.Y., Braden, B.B., Byrge, L., Castellanos, F.X., Dapretto, M., Delorme, R., Fair, D.A., Fishman, I., Fitzgerald, J., Gallagher, L., Keehn, R.J.J., Kennedy, D.P., Lainhart, J.E., Luna, B., Mostofsky, S.H., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Solomon, M., Toro, R., Vaidya, C.J., Wenderoth, N., White, T., Craddock, R.C., Lord, C., Leventhal, B., Milham, M.P., 2017. Data Descriptor: Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Scientific Data 4.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner,

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Muller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19, 659-667.

Dice, L.R., 1945. Measures of the Amount of Ecologic Association between Species. Ecology 26, 297-302.

Finn, E.S., Shen, X.L., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nature Neuroscience 18, 1664-1671.

Glover, G.H., Mueller, B.A., Turner, J.A., van Erp, T.G., Liu, T.T., Greve, D.N., Voyvodic, J.T., Rasmussen, J., Brown, G.G., Keator, D.B., Calhoun, V.D., Lee, H.J., Ford, J.M., Mathalon, D.H., Diaz, M., O'Leary, D.S., Gadde, S., Preda, A., Lim, K.O., Wible, C.G., Stern, H.S., Belger, A., McCarthy, G., Ozyurt, B., Potkin, S.G., 2012. Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J Magn Reson Imaging 36, 39-54.

Gotts, S.J., Saad, Z.S., Jo, H.J., Wallace, G.L., Cox, R.W., Martin, A., 2013. The perils of global signal regression for group comparisons: a case study of Autism Spectrum Disorders. Front Hum Neurosci 7, 356.

Gountouna, V.E., Job, D.E., McIntosh, A.M., Moorhead, T.W., Lymer, G.K., Whalley, H.C., Hall, J., Waiter, G.D., Brennan, D., McGonigle, D.J., Ahearn, T.S., Cavanagh, J., Condon, B., Hadley, D.M., Marshall, I., Murray, A.D., Steele, J.D., Wardlaw, J.M., Lawrie, S.M., 2010. Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49, 552-560.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proc Natl Acad Sci U S A 100, 253-258.

He, Y., Lim, S., Fortunato, S., Sporns, O., Zhang, L., Qiu, J., Xie, P., Zuo, X.N., 2018. Reconfiguration of Cortical Networks in MDD Uncovered by Multiscale Community Detection with fMRI. Cerebral Cortex 28, 1383-1395.

Holiga, S., Hipp, J.F., Chatham, C.H., Garces, P., Spooren, W., D'Ardhuy, X.L., Bertolino, A., Bouquet, C., Buitelaar, J.K., Bours, C., Rausch, A., Oldehinkel, M., Bouvard, M., Amestoy, A., Caralp, M., Gueguen, S., Ly-Le Moal, M., Houenou, J., Beckmann, C.F., Loth, E., Murphy, D., Charman, T., Tillmann, J., Laidi, C., Delorme, R., Beggiato, A., Gaman, A., Scheid, I., Leboyer, M., d'Albis, M.A., Sevigny, J., Czech, C., Bolognani, F., Honey, G.D., Dukart, J., 2019. Patients with autism spectrum disorders display reproducible functional connectivity alterations. Science Translational Medicine 11.

Hull, J.V., Dokovna, L.B., Jacokes, Z.J., Torgerson, C.M., Irimia, A., Van Horn, J.D., 2016. Resting-State Functional Connectivity in Autism Spectrum Disorders: A Review. Front Psychiatry 7, 205.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825-841.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Jones, T.B., Bandettini, P.A., Kenworthy, L., Case, L.K., Milleville, S.C., Martin, A., Birn, R.M., 2010. Sources of group differences in functional connectivity: an investigation applied to autism spectrum disorder. Neuroimage 49, 401-414.

King, J.B., Prigge, M.B.D., King, C.K., Morgan, J., Dean, D.C., Freeman, A., Villaruz, J.A.M., Kane, K.L., Bigler, E.D., Alexander, A.L., Lange, N., Zielinski, B.A., Lainhart, J.E., Anderson, J.S., 2018. Evaluation of Differences in Temporal Synchrony Between Brain Regions in Individuals With Autism and Typical Development. Jama Network Open 1.

Lemieux, L., Salek-Haddadi, A., Lund, T.E., Laufs, H., Carmichael, D., 2007. Modelling large motion events in fMRI studies of patients with epilepsy. Magn Reson Imaging 25, 894-901.

Müller, R.-A., Shih, P., Keehn, B., Deyoe, J.R., Leyden, K.M., Shukla, D.K., 2011. Underconnected, but How? A Survey of Functional Connectivity MRI Studies in Autism Spectrum Disorders. Cerebral Cortex 21, 2233-2243.

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.

Nair, S., Jao Keehn, R.J., Berkebile, M.M., Maximo, J.O., Witkowska, N., Muller, R.A., 2018. Local resting state functional connectivity in autism: site and cohort variability and the effect of eye status. Brain Imaging Behav 12, 168-179.

Nee, D.E., 2019. fMRI replicability depends upon sufficient individual-level data. Communications Biology 2, 130.

Noble, S., Scheinost, D., Finn, E.S., Shen, X., Papademetris, X., McEwen, S.C., Bearden, C.E., Addington, J., Goodyear, B., Cadenhead, K.S., Mirzakhanian, H., Cornblatt, B.A., Olvet, D.M., Mathalon, D.H., McGlashan, T.H., Perkins, D.O., Belger, A., Seidman, L.J., Thermenos, H., Tsuang, M.T., van Erp, T.G.M., Walker, E.F., Hamann, S., Woods, S.W., Cannon, T.D., Constable, R.T., 2017. Multisite reliability of MR-based functional connectivity. Neuroimage 146, 959-970.

Parkes, L., Fulcher, B., Yucel, 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.

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.

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.

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.

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.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52, 1059-1069.

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

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.

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.

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.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.N., Holmes, A.J., Eickhoff, S.B., Thomas, B.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex 28, 3095-3114.

Scholvinck, 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, 10238-10243.

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y.Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23, S208-S219.

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E., 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. Nature Neuroscience 18, 1565-1567.

Turner, J.A., Damaraju, E., van Erp, T.G., Mathalon, D.H., Ford, J.M., Voyvodic, J., Mueller, B.A., Belger, A., Bustillo, J., McEwen, S., Potkin, S.G., Fbirn, Calhoun, V.D., 2013. A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia. Front Neurosci 7, 137.

Tyszka, J.M., Kennedy, D.P., Paul, L.K., Adolphs, R., 2014. Largely typical patterns of resting-state functional connectivity in high-functioning adults with autism. Cerebral Cortex 24, 1894-1905.

Uddin, L.Q., 2017. Mixed Signals: On Separating Brain Signal from Noise. Trends Cogn Sci 21, 405-406.

Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L., 2012. The influence of head motion on intrinsic functional connectivity MRI. Neuroimage 59, 431-438.

Varikuti, D.P., Hoffstaedter, F., Genon, S., Schwender, H., Reid, A.T., Eickhoff, S.B., 2017. Resting-state test-retest reliability of a priori defined canonical networks over different preprocessing steps. Brain Struct Funct 222, 1447-1468.

Yahata, N., Morimoto, J., Hashimoto, R., Lisi, G., Shibata, K., Kawakubo, Y., Kuwabara, H., Kuroda, M., Yamada, T., Megumi, F., Imamizu, H., Nanez, J.E., Sr., Takahashi, H., Okamoto, Y., Kasai, K., Kato, N., Sasaki, Y., Watanabe, T., Kawato, M., 2016. A small number of abnormal brain connections predicts adult autism spectrum disorder. Nat Commun 7, 11254.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Yamagata, H., Matsuo, K., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Kasai, K., Kato, N., Takahashi, H.,

bioRxiv preprint first posted online May. 17, 2019; doi: http://dx.doi.org/10.1101/640797. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Okamoto, Y., Tanaka, S.C., Kawato, M., Yamashita, O., Imamizu, H., 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLOS Biology 17, e3000042.

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

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

Yang, G.J., Murray, J.D., Repovs, G., Cole, M.W., Savic, A., Glasser, M.F., Pittenger, C., Krystal, J.H., Wang, X.J., Pearlson, G.D., Glahn, D.C., Anticevic, A., 2014. Altered global brain signal in schizophrenia. Proc Natl Acad Sci U S A 111, 7438-7443.

Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J Neurophysiol 106, 1125-1165.

Yu, M., Linn, K.A., Cook, P.A., Phillips, M.L., McInnis, M., Fava, M., Trivedi, M.H., Weissman, M.M., Shinohara, R.T., Sheline, Y.I., 2018. Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data. Hum Brain Mapp 39, 4213-4227.