




RESEARCH ARTICLE

Test–retest reproducibility of white matter parcellation using diffusion MRI tractography fiber clustering

Fan Zhang  | Ye Wu | Isaiah Norton | Yogesh Rathi  | Alexandra J. Golby |
Lauren J. O'Donnell 

Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Correspondence

Fan Zhang, Brigham and Women's Hospital, Harvard Medical School, 25 Shattuck St, Boston, MA 02115.

Email: fzhang@bwh.harvard.edu

Funding information

National Institutes of Health, Grant/Award Numbers: R03 NS088301, U01 CA199459, R01 MH097979, R01 MH074794, P41 EB015898, P41 EB015902

Abstract

There are two popular approaches for automated white matter parcellation using diffusion MRI tractography, including fiber clustering strategies that group white matter fibers according to their geometric trajectories and cortical-parcellation-based strategies that focus on the structural connectivity among different brain regions of interest. While multiple studies have assessed test–retest reproducibility of automated white matter parcellations using cortical-parcellation-based strategies, there are no existing studies of test–retest reproducibility of fiber clustering parcellation. In this work, we perform what we believe is the first study of fiber clustering white matter parcellation test–retest reproducibility. The assessment is performed on three test–retest diffusion MRI datasets including a total of 255 subjects across genders, a broad age range (5–82 years), health conditions (autism, Parkinson's disease and healthy subjects), and imaging acquisition protocols (three different sites). A comprehensive evaluation is conducted for a fiber clustering method that leverages an anatomically curated fiber clustering white matter atlas, with comparison to a popular cortical-parcellation-based method. The two methods are compared for the two main white matter parcellation applications of dividing the entire white matter into parcels (i.e., whole brain white matter parcellation) and identifying particular anatomical fiber tracts (i.e., anatomical fiber tract parcellation). Test–retest reproducibility is measured using both geometric and diffusion features, including volumetric overlap (wDice) and relative difference of fractional anisotropy. Our experimental results in general indicate that the fiber clustering method produced more reproducible white matter parcellations than the cortical-parcellation-based method.

KEYWORDS

cortical-parcellation-based, diffusion MRI, fiber clustering, tractography parcellation

1 | INTRODUCTION

Diffusion magnetic resonance imaging (dMRI) provides the only existing technique to map the structural connections of the living human brain in a noninvasive way (Basser, Mattiello, & LeBihan, 1994). dMRI allows the estimation of white matter fiber tracts in the brain via a process called tractography (Basser, Pajevic, Pierpaoli, Duda, & Aldroubi, 2000), which has been widely used for understanding neurological development, brain function, and brain disease, as described in several reviews (Ciccarelli, Catani, Johansen-Berg, Clark, & Thompson, 2008; Essayed et al., 2017; Pannek, Scheck, Colditz, Boyd, & Rose, 2014; Piper, Yoong, Kandasamy, & Chin, 2014; Yamada, Sakai, Akazawa,

Yuen, & Nishimura, 2009). White matter parcellation, that is, dividing the massive number of tractography fibers (streamline trajectories) into multiple fiber parcels (or fiber tracts), is the first and essential step to enable fiber quantification and visualization. White matter parcellation can enable the study of *fiber parcels from the entire white matter* to identify between-population differences (e.g., between patients harboring disease and healthy subjects) using machine learning or statistical analyses (Ingalhalikar et al., 2014; Sporns, Tononi, & Kötter, 2005; Zalesky, Cocchi, Fornito, Murray, & Bullmore, 2012; Zhang, Savadjiev, et al., 2018; Zhang, Wu, Ning, 2018). White matter parcellation is also important for identifying *anatomical fiber tracts* for clinical visualization (Golby et al., 2011; Gong et al., 2018; Nimsky,

Ganslandt, Dorit, Gregory Sorensen, & Fahlbusch, 2006; O'Donnell et al., 2017) or hypothesis-driven research (Alexander et al., 2007; Shany et al., 2017; Wu et al., 2015, 2018; Yeo, Jang, & Son, 2014). Automated and robust white matter parcellation can enable the analysis of new, large dMRI datasets that are being acquired to study complex neural systems across the lifespan and across brain disorders (Alexander et al., 2017; Casey et al., 2018; Thompson et al., 2017).

There are two popular strategies for automated white matter parcellation (O'Donnell, Golby, & Westin, 2013): (a) *fiber clustering* that groups white matter fibers according to their geometric trajectories, aiming to reconstruct tracts corresponding to the white matter anatomy (Ding, Gore, & Anderson, 2003; Garyfallidis et al., 2018; Garyfallidis, Brett, Correia, Williams, & Nimmo-Smith, 2012; Guevara et al., 2012; Jin et al., 2014; Kumar, Desrosiers, Siddiqi, Colliot, & Toews, 2017; O'Donnell & Westin, 2007; Prasad et al., 2014; Siless, Chang, Fischl, & Yendiki, 2018; Visser, Nijhuis, Buitelaar, & Zwiers, 2011; Wassermann, Bloy, Kanterakis, Verma, & Deriche, 2010; Zhang, Wu, Norton, 2018) and (b) *cortical-parcellation-based* that parcellates tractography according to a cortical parcellation, focusing on the structural connectivity among different brain regions of interest (ROIs) (Bassett & Bullmore, 2016; Bastiani, Shah, Goebel, & Roebroeck, 2012; Bullmore & Sporns, 2009; Gong et al., 2009; Ingalhalikar et al., 2014; Sporns et al., 2005; Wakana et al., 2007; Wassermann et al., 2016; Yeh, Badre, & Verstynen, 2016; Zalesky et al., 2012; Zhang et al., 2010).

An important goal of white matter parcellation is to identify white matter structures that are reproducible (O'Donnell & Pasternak, 2015). *Test-retest reproducibility* assesses whether parcellated white matter structures can be reliably reproduced for the same individual subject in repeated (test-retest) dMRI scans. Test-retest reproducibility is considered to be a good indicator of the reliability of white matter parcellation for potential clinical applications (Besseling et al., 2012; Jovicich et al., 2014; Keihaninejad et al., 2013; Kristo et al., 2013; Lin et al., 2013). To measure test-retest reproducibility, previous works have used *geometrical measures* such as volume, volumetric overlap, fiber length and shape, and number of fibers (Cheng et al., 2012; Cousineau et al., 2017; Kristo et al., 2013; Lin et al., 2013; Owen, Chang, & Mukherjee, 2015; Smith, Tournier, Calamante, & Connelly, 2015; Wang et al., 2012; Zhao et al., 2015). Other groups have used *diffusion measures* such as fractional anisotropy (FA) and mean diffusivity (MD) computed from the voxels through which the parcellated fibers pass (Besseling et al., 2012; Ciccarelli et al., 2003; Duan, Zhao, He, & Shu, 2015; Kristo et al., 2013; Papinutto, Maule, & Jovicich, 2013; Pfefferbaum, Adalsteinsson, & Sullivan, 2003; Vollmar et al., 2010; Yendiki, Reuter, Paul, Diana Rosas, & Fischl, 2016). While multiple studies have assessed test-retest reproducibility of white matter parcellations using cortical-parcellation-based strategies, for example, on the brain connectome network (Besson, Lopes, Leclerc, Derambure, & Tyvaert, 2014; Bonilha et al., 2015; Buchanan, Pernet, Gorgolewski, Storkey, & Bastin, 2014; Dennis et al., 2012; Duda, Cook, & Gee, 2014; Schumacher et al., 2018; Smith et al., 2015; Vaessen et al., 2010; Zhao et al., 2015; Zhang, Descoteaux, et al., 2018) and on anatomical fiber tracts (Besseling et al., 2012; Cousineau et al., 2017; Heiervang, Behrens, Mackay, Robson, & Johansen-Berg, 2006; Kristo et al., 2013; Lin et al., 2013; Papinutto et al., 2013; Tensaouti, Lahlou, Clarisse, Lotterie, & Berry, 2011; Wang et al., 2012; Yendiki et al., 2016), there are no existing

studies of fiber clustering, to our knowledge. Studies have suggested that fiber clustering approaches have advantages in parcellating the white matter in a highly consistent way, aiming to reconstruct fiber parcels/tracts corresponding to the white matter anatomy (Ge et al., 2012; Sydnor et al., 2018; Zhang et al., 2017; Zhang, Wu, Norton, et al., 2018; Ziyang, Sabuncu, Eric, Grimson, & Westin, 2009), while cortical-parcellation-based methods could be less consistent considering factors such as the variability of intersubject cortical anatomy, the dependence on registration between dMRI and structural images, and the presence of false positive/negative connections (Amunts et al., 1999; Fischl, Sereno, Tootell, & Dale, 1999; Maier-Hein et al., 2017; Sinke et al., 2018; Zhang, Descoteaux, et al., 2018). However, to our knowledge, test-retest reproducibility of the cortical-parcellation-based and fiber clustering white matter parcellation strategies has not yet been quantitatively compared.

In the present work, we conduct what we believe is the first study to investigate the test-retest reproducibility of fiber clustering white matter parcellation. An fiber clustering method based on an anatomically curated white matter fiber clustering atlas (Zhang, Wu, Norton, et al., 2018) is evaluated, with comparison to a cortical-parcellation-based method based on a cortical and subcortical parcellation from Freesurfer (Fischl, 2012). Both of these white matter parcellation methods have been demonstrated to be successful in comparison to manual tract parcellation (O'Donnell et al., 2017; Sydnor et al., 2018; Wassermann et al., 2016). Here, the test-retest reproducibility of the two methods is compared for two main applications: (a) *whole brain white matter parcellation*, that is, dividing the entire white matter into fiber parcels and (b) *anatomical fiber tract parcellation*, that is, identifying particular anatomical fiber tracts. Multiple quantitative measurements, including a geometrical measure (volumetric overlap) and a diffusion measure (FA), are computed for test-retest reproducibility evaluation. A large test-retest dataset is studied, including a total of 255 subjects from multiple independently acquired populations.

2 | MATERIALS AND METHODS

2.1 | Datasets, data processing, and tractography

In this study, we evaluated the test-retest reproducibility of two white matter parcellation methods on dMRI data from a total of 255 subjects across genders (87 females vs. 168 males), a broad age range (children, young adults, and older adults, from 5 to 82 years), and different health conditions (autism, Parkinson's disease, and healthy subjects). This publicly available test-retest dMRI data were from three independently acquired datasets with different diffusion imaging protocols. Table 1 gives an overview of the datasets studied, including demographic information and diffusion image acquisitions.

For each of the subjects under study, the test-retest dMRI scans were preprocessed to exclude any potential artifacts, for example, from eddy current and head motion effects. Details of the preprocessing steps for each dataset under study are included in Appendix. Whole brain tractography was then computed using the two-tensor unscented Kalman filter (UKF) method (Malcolm, Shenton, & Rathi, 2010; Reddy & Rathi, 2016), as implemented in the *ukftractography* package (<https://github.com/pnlbwh/ukftractography>). The UKF method fits a mixture

TABLE 1 Demographics and dMRI data of the datasets under study

Dataset	# subjects	Age	Gender	Health condition	dMRI data studied
ABIDE II	70	5–17 years (12.0 ± 3.1)	6 F 64 M	49 AUT 21 healthy	1 b0 image 63 gradient directions ($b = 1,000$) TE/TR = 78/5200 ms Resolution = 3 mm ³ (isotropic) Test–retest interval: In same scan session
HCP	44	22–35 years (30.4 ± 3.3)	31 F 13 M	44 healthy	18 b0 images 90 gradient directions ($b = 3,000$) TE/TR = 89/5520 ms Resolution = 1.25 mm ³ (isotropic) Test–retest interval: 18–328 days (134 ± 62)
PPMI	141	51–82 years (63.7 ± 7.2)	50 F 91 M	100 PD 41 healthy	1 b0 image 63 gradient directions ($b = 1,000$) TE/TR = 88/7600 ms Resolution = 2 mm ³ (isotropic) Test–retest interval: In same scan session

Abbreviations. Dataset—ABIDE-II: autism brain imaging data exchange II (Martino et al., 2017); HCP: human connectome project (Van Essen et al., 2013); PPMI: Parkinson's Progression Markers Initiative (Marek et al., 2011); Disease—AUT: autism; PD: Parkinson's disease. Gender—F: female; M: male.

model of two tensors to the dMRI data while tracking fibers, employing prior information from the previous step to help stabilize model fitting. We chose the UKF method because it has been shown to be highly consistent in tracking fibers in dMRI data from independently acquired populations across ages, health conditions and image acquisitions (Zhang, Wu, Norton, et al., 2018), and it is more sensitive than standard single-tensor tractography, in particular in the presence of crossing fibers and peritumoral edema (Baumgartner et al., 2012; Chen et al., 2015, 2016; Liao et al., 2017). For each of the subjects under study, the tractography produced about 1 million fibers per dMRI scan. Details of the related tractography parameters are included in Appendix.

For each subject, the tractography datasets computed from the two test–retest dMRI scans were aligned into the same space. This was conducted by computing a registration between the FA images computed from the dMRI scans. The Advanced Normalization Tools (ANTs) package (<https://github.com/ANTsX/ANTs>) (Avants, Tustison, & Song, 2009) was used to perform the registration, following the default steps in the software including a rigid, an affine, then a deformable transformation to reach a good intra-subject alignment. Here, we chose an FA-based registration because the FA image is sensitive to white matter fiber tracts and can provide a good correspondence between the registered fiber tracts (Goodlett, Davis, Jean, Gilmore, & Gerig, 2006), and it has been applied in many studies for registering between dMRI datasets (Besseling et al., 2012; Papinutto et al., 2013; Vollmar et al., 2010). In details, this registration was performed by aligning the FA image from the second scan to that from the first scan. The obtained transforms were then applied to the tractography data computed from the second scan, using 3D Slicer. This step ensured the tractography data from the second scan was registered to the space of the first scan. In this way, because transforms were applied after performing tractography, any resampling or blurring of diffusion-weighted image data was avoided.

2.2 | White matter parcellation

After obtaining whole brain tractography, white matter parcellation was performed using the two methods (fiber clustering and cortical-parcellation-based), as illustrated in Figure 1. For each method, both

whole brain white matter parcellation and anatomical fiber tract parcellation were performed. Details of each parcellation method are introduced below.

2.2.1 | Fiber clustering white matter parcellation

The fiber clustering method performs white matter parcellation of an individual subject based on an fiber clustering atlas (Figure 1a1.) provided by the O'Donnell Research Group (ORG) (<http://dmri.slicer.org/atlas>) (Zhang, Wu, Norton, et al., 2018). The ORG atlas contains an 800-cluster parcellation of the entire white matter and an anatomical fiber tract parcellation, including 58 deep white matter fiber tracts, plus 198 short and medium range superficial fiber clusters organized into 16 categories according to the brain lobes they connect. (The atlas was generated by creating dense tractography maps [using the same UKF tractography method as in the current study]) of 100 individual HCP subjects and then applying an fiber clustering method to group the tracts across subjects according to their similarity in shape and location. The resulting clusters were annotated using expert neuroanatomical knowledge.) We chose the ORG-atlas-based fiber clustering method because it is an anatomically curated white matter atlas that has been demonstrated to have a good performance for consistent white matter parcellation across different populations (Zhang, Descoteaux, et al., 2018).

The fiber clustering method was applied to perform white matter parcellation of one subject as follows. A tractography-based registration was performed to align the subject's tractography data into the atlas space. A fiber spectral embedding was conducted to compute the similarity of fibers between the subject and the atlas, followed by the assignment of each fiber of the subject to the corresponding atlas cluster. This process produced a whole brain white matter parcellation into 800 fiber clusters. (This parcellation scale of 800 fiber clusters has been chosen to successfully separate white matter structures considered to be anatomically different (O'Donnell et al., 2017; Zhang, Wu, Ning, et al., 2018; Zhang, Wu, Norton, et al., 2018). These fiber clusters included 84 commissural clusters as well as 716 bilateral hemispheric clusters (that included fibers in both hemispheres). We separated the hemispheric clusters by hemisphere (the maximum number of clusters is thus $[716 \times 2 + 84] = 1,516$); therefore, we

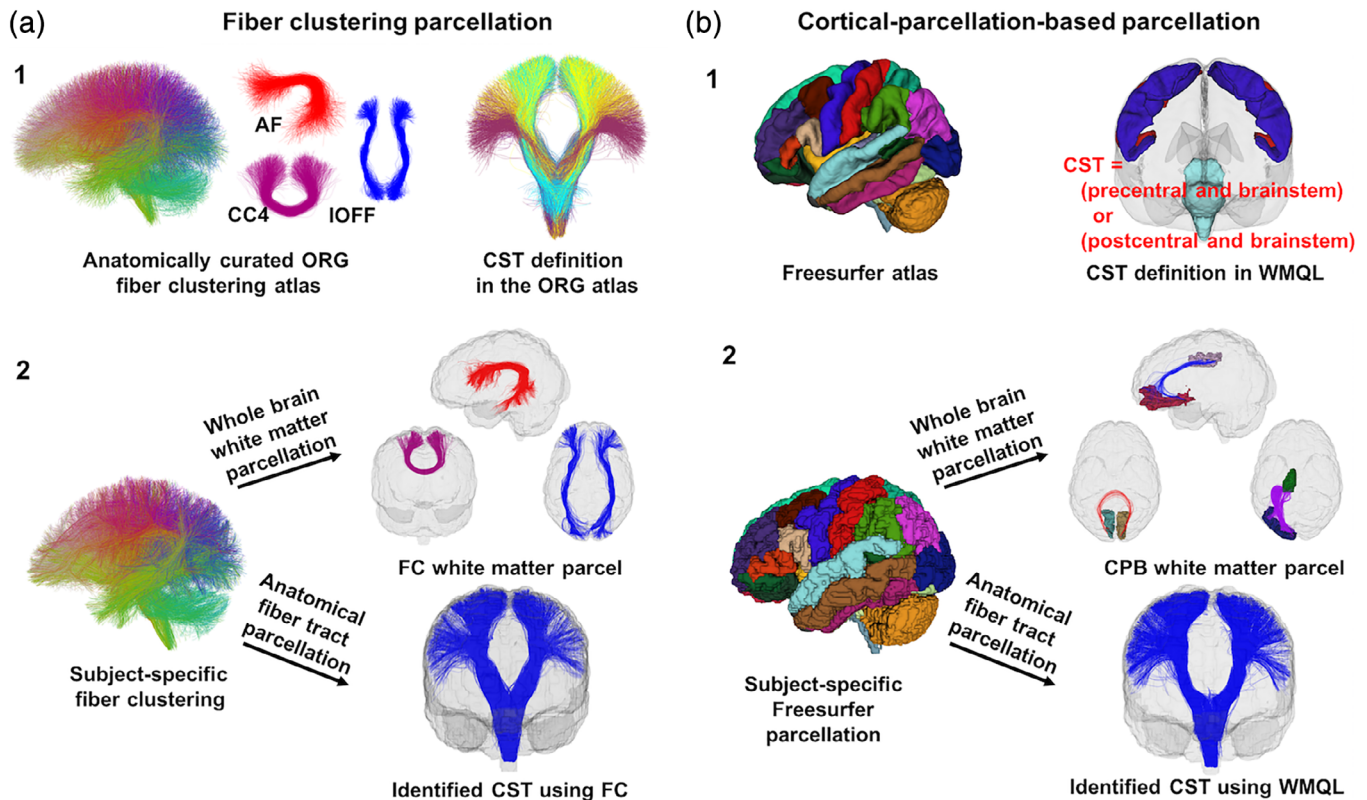


FIGURE 1 Overview of the two white matter parcellation methods. Sub-figure (a) shows the fiber clustering (FC) method. It relies on an O'Donnell's Research Group (ORG) fiber clustering atlas (a1) that includes an 800-cluster parcellation of the entire white matter and an anatomical fiber tract parcellation. Whole brain white matter parcellation (a2) is performed by identifying subject-specific fiber clusters according to the 800-cluster atlas parcellation (a2). Anatomical fiber tract parcellation (a2) is performed by leveraging the anatomically curated tracts in the atlas. Sub-figure (b) shows the cortical-parcellation-based (CPB) method. It relies on a neuroanatomical brain parcellation atlas from Freesurfer to segment an individual's brain into multiple cortical and subcortical regions (b1). Whole brain white matter parcellation (b2) is performed by identifying fiber parcels connecting between each pair of the segmented regions of interest (ROIs). Anatomical fiber tract parcellation (b2) is performed by leveraging White matter query language (WMQL), which provides anatomical definitions of fiber tracts based on their intersected Freesurfer regions (e.g., the CST) [Color figure can be viewed at wileyonlinelibrary.com]

produced over 800 parcels for each subject (see experimental results in Section 3.1 for the number of parcels). Next, for anatomical fiber tract parcellation, we leveraged the anatomically curated tracts in the atlas. Each tract was comprised of a set of fiber clusters. In our study, we used 45 tracts (see Table 2 for the list of the tracts) that are defined in both the fiber clustering method and the cortical-parcellation-based method (Section 2.2.2). All fiber clustering processing was performed using the *whitematteranalysis* software (<https://github.com/SlicerDMRI/whitematteranalysis>), and all parameters were set to their default values. (Details are described in (O'Donnell et al., 2017; O'Donnell, Wells, Golby, & Westin, 2012; O'Donnell & Westin, 2007; Zhang, Wu, Norton, et al., 2018).

2.2.2 | Cortical-parcellation-based white matter parcellation

The cortical-parcellation-based method performs white matter parcellation based on Freesurfer (<http://freesurfer.net>; Fischl, 2012). Freesurfer anatomically segments brain regions of an individual subject based on brain atlases including a cortical (Desikan et al., 2006) and a subcortical parcellation (Fischl et al., 2002; Figure 1b1). We chose the Freesurfer-based cortical-parcellation-based method for comparison because it has been applied in many studies of the test-

retest reproducibility of whole brain white matter parcellation (Besson et al., 2014; Bonilha et al., 2015; Buchanan et al., 2014; Smith et al., 2015) and anatomical fiber tract parcellation (Cousineau et al., 2017; Ning et al., 2016; Roy et al., 2017; Yendiki et al., 2016).

The cortical-parcellation-based method was applied to perform white matter parcellation of one subject as follows. The Freesurfer segmentation of an individual subject (Figure 1b2) was performed using the T1-weighted image. Registration between the T1-weighted and dMRI images was performed (see Appendix) so that the tractography data was in the same space as the Freesurfer parcellation result. Then, for whole brain white matter parcellation, we identified parcels connecting between all pairs of segmented cortical and subcortical regions (Figure 1b2), as in many previous studies (Besson et al., 2014; Bonilha et al., 2015; Buchanan et al., 2014; Smith et al., 2015). The Freesurfer atlases (Desikan et al., 2006; Fischl et al., 2002) include a total of 87 ROIs. Therefore, the cortical-parcellation-based whole brain parcellation resulted in a total of 3,741 ($87 \times 86/2$) parcels per subject. Next, for anatomical fiber tract parcellation, we leveraged the White Matter Query Language (WMQL; https://github.com/demianw/tract_querier; Wassermann et al., 2016), an automated method to delineate anatomical fiber tracts based on the Freesurfer anatomical regions they intersect (Figure 1b1). In our study, we applied WMQL because it

enables identification of a relatively large number of fiber tracts (45 tracts) and it has been used in multiple works to study white matter parcellation retest–retest reproducibility (Cousineau et al., 2017; Ning et al., 2016; Roy et al., 2017). For a given anatomical fiber tract, subject-specific parcellation was performed by identifying the fibers from the whole brain tractography that met the tract's anatomical definition (Figure 1b2). In our work, a total of 45 anatomical fiber tracts that had available WMQL tract definitions were extracted for each subject under study (the same tracts as in the fiber clustering method; see Table 2 for the list of the tracts).

2.3 | Test–retest measurements

After performing white matter parcellation, we computed test–retest measurements of the parcellated white matter structures (whole-brain fiber parcels or anatomical fiber tracts) to evaluate the reproducibility of the parcellation performance. We included both geometrical and diffusion measures.

For a geometrical measure, we computed the volumetric overlap between the parcellated white matter structures to investigate if they had the same volume and shape. We applied the weighted Dice (wDice) coefficient that was designed specifically for measuring volumetric overlap of fiber tracts (Cousineau et al., 2017). wDice extends the standard Dice coefficient (Dice, 1945) taking account of the number of fibers per voxel so that it gives higher weighting to voxels with dense fibers, as follows:

TABLE 2 A total of 45 fiber tracts are compared between the fiber clustering and cortical-parcellation-based methods, including 24 hemispheric (LR) association tracts, 7 commissural (C) tracts, and 14 hemispheric (LR) projection tracts

Tract category (number of tracts)	Tract name
Association tracts (24)	Arcuate fasciculus (AF)—LR
	Cingulum bundle (CB)—LR
	External capsule (EC)—LR
	Extreme capsule (EmC)—LR
	Inferior longitudinal fasciculus (ILF)—LR
	Inferior occipito-frontal fasciculus (IoFF)—LR
	Middle longitudinal fasciculus (MdLF)—LR
	Posterior limb of internal capsule (PLIC)—LR
	Superior longitudinal fasciculus I (SLF I)—LR
	Superior longitudinal fasciculus II (SLF II)—LR
	Superior longitudinal fasciculus II (SLF III)—LR
	Uncinate fasciculus (UF)—LR
	Commissural tracts (7)
Corpus callosum 2 (CC 2)—C	
Corpus callosum 3 (CC 3)—C	
Corpus callosum 4 (CC 4)—C	
Corpus callosum 5 (CC 5)—C	
Corpus callosum 6 (CC 6)—C	
Corpus callosum 7 (CC 7)—C	
Projection tracts (14)	Corticospinal tract (CST)—LR
	Striato-frontal (SF)—LR
	Striato-occipital (SO)—LR
	Striato-parietal (SP)—LR
	Thalamo-frontal (TF)—LR
	Thalamo-occipital (TO)—LR
Thalamo-parietal (TP)—LR	

These tracts are the ones that can be identified in both of the parcellation methods.

$$wDice(P1, P2) = \frac{\sum_{v'} W_{1,v'} + \sum_{v'} W_{2,v'}}{\sum_{v'} W_{1,v} + \sum_{v'} W_{2,v}} \quad (1)$$

where $P1$ and $P2$ represent two corresponding parcellated white matter structures from the test–retest data, v' indicates the set of voxels that are within the intersection of the volumes of $P1$ and $P2$, v indicates the set of voxels that are within the union of the volumes of $P1$ and $P2$, and W is the fraction of the fibers passing through a voxel. A high wDice value represents a high reproducibility between the two corresponding parcellated white matter structures.

Then, for a diffusion measure, we calculated the reproducibility of the mean FA of the voxels where the parcellated white matter structures were located. In related work, Papinutto et al. evaluated reproducibility by computing the absolute difference between mean FA values divided by the average of the two mean FA values (Papinutto et al., 2013). In our study, we adopted a similar evaluation strategy and extended it by measuring a relative difference, that is, the absolute difference divided by the sum of the mean FA values. We chose this approach because diffusion properties (e.g., FA) are different across different white matter structures (Madden et al., 2004; Piepaoli, Jezzard, Basser, Barnett, & Di Chiro, 1996; Santis, Silvia, Bells, Assaf, & Jones, 2014), and relative difference can provide comparable values across different structures. This was essential for comparing the reproducibility of the whole brain parcellations across methods, because there was no one-to-one correspondence between the white matter parcels produced by the two methods. Specifically, for two corresponding parcellated white matter structures $P1$ and $P2$ from the test–retest scans, we computed the mean FA of the voxels where their fibers passed, and measured the relative difference of the mean FA values, as follows:

$$RD(P1, P2) = \left| \frac{FA(P1) - FA(P2)}{FA(P1) + FA(P2)} \right| \quad (2)$$

A low relative difference value represents a high reproducibility between the two parcellated white matter structures. (We note that we have provided the results of the reproducibility of the mean MD, analyzed in the same way as the mean FA, in Supporting Information S1.)

2.4 | Statistical analysis

We then performed statistical comparisons between the two parcellation methods based on the computed test–retest measurements. These statistics were performed in each of the three datasets.

2.4.1 | Whole brain white matter parcellation analysis

The parcellation methods were compared in two ways, including a comparison using all parcels and a more fine-grained comparison using parcels with similar volumes. First, for each parcel, we computed the mean wDice score and mean relative difference value across all subjects for each parcellation method. This gave a vector of mean wDice scores (and another vector of mean relative difference values), with length of the number of parcels, for each method. We compared the wDice (and also relative difference) measurements across methods using an unpaired two-sample Student's (two-tailed) t -test, for each of the three datasets separately. Second, we compared parcels with similar volumes across the two methods. To enable this, we created a

histogram by binning parcels into 50 bins evenly distributed between 0 and 100,000 mm³ (approximately the maximum volume size across all parcels). To compare parcels in each bin across methods, we performed an unpaired two-sample Student's (two-tailed) *t*-test to analyze both the mean wDice and mean relative difference of these parcels. The false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995) was used to control for multiple comparisons (across all bins). We excluded a bin if the number of parcels in either method was 0. We note that, to eliminate potential biases from inconsistent parcels, for each method we included only the parcels that were detected in at least 95% of subjects (see Table 3 for the number of retained parcels).

2.4.2 | Anatomical fiber tract parcellation analysis

The parcellation methods were compared in two ways, including a comparison using all tracts and a more fine-grained comparison using individual tracts. First, for each tract, we computed the mean wDice and mean relative difference across all subjects for each parcellation method. We compared these measurements between methods using a paired two-sample Student's (two-tailed) *t*-test. (The paired test was performed because the same set of anatomical tracts was selected in the fiber clustering method to match the set of tracts parcellated in the cortical-parcellation-based method.) Second, we compared individual tracts. For each tract, we analyzed the wDice and relative difference from all subjects using a paired two-sample Student's (two-tailed) *t*-test. The FDR procedure was used to control for multiple comparisons across all tracts.

3 | RESULTS

3.1 | Whole brain white matter parcellation results

Table 3 shows the number and volume of retained white matter parcels after removing the ones that were not consistently detected across each dataset (Section 2.4.1). While the number of retained parcels was different between the two methods across the three datasets ($p = 0.006$; paired *t*-test, two-tailed), their median parcel volumes were similar ($p = 0.151$; paired *t*-test, two-tailed).

3.1.1 | Volumetric overlap of whole brain white matter parcels

Figure 2 shows the distributions of the mean wDice scores in each dataset, after application of the fiber clustering and cortical-parcellation-based methods. Because there was no one-to-one correspondence of the parcels between the two methods, we plotted the mean wDice of each parcel versus its mean volume. This enables a

visual comparison of method performance for parcels with various volumes. It is visually apparent that the mean wDice was generally lower and had a larger range in the cortical-parcellation-based method than the fiber clustering method. For quantitative assessment, in the comparison using all parcels, significantly higher mean wDice was achieved using the fiber clustering method compared to the cortical-parcellation-based method, with $p < 0.001$ (unpaired *t*-test, two-tailed) for each of the three datasets. In the more fine-grained comparison using parcels with similar volumes (parcels within the same bin of the parcel volume histogram), 63.16%, 58.97%, and 69.70% of all retained bins, respectively, in the ABIDE-II, HCP and PPMI datasets, had significantly higher mean wDice scores using the fiber clustering method ($p < 0.05$, unpaired *t*-test, two-tailed; FDR corrected), while no bins had significantly higher mean wDice scores using the cortical-parcellation-based method. (See Supporting Information 2 for the number of the retained bins in each dataset after removing the ones that had 0 parcels in either method.)

For visual quality assessment of the volumetric overlap performance, Figure 3 gives a visual comparison of the white matter parcels obtained using the fiber clustering and cortical-parcellation-based methods. We identified the most and the least reproducible parcels from both methods, in terms of wDice score, for a certain parcel volume. In Figure 3, white matter parcels with volume around 10,000 mm³ (from one subject in the HCP dataset) are displayed. This volume was chosen as approximately the average of the median of the parcel volumes of the six parcellations (two parcellation methods for each of the three datasets).

3.1.2 | Relative difference of FA of whole brain white matter parcels

Figure 4 shows the distributions of the mean relative difference of the white matter parcels in the fiber clustering and cortical-parcellation-based methods, versus the mean parcel volume, in each dataset. We can observe that fiber clustering method obtained visually lower mean relative difference than the cortical-parcellation-based method; however, this visual difference is not as apparent as that observed in the mean wDice (Figure 2). For quantitative assessment, in the comparison using all parcels, significantly lower mean relative difference was achieved using the fiber clustering method compared to the cortical-parcellation-based method, with $p < 0.001$ (unpaired *t*-test, two-tailed) for each of the three datasets. In the more fine-grained comparison using parcels with similar volumes (parcels within the same bin of the parcel volume histogram), 7.89%, 23.08%, and 48.48% of all retained bins, respectively, in the ABIDE-II, HCP and PPMI datasets, had significantly lower mean relative difference values using the fiber clustering method ($p < 0.05$, unpaired *t*-test, two-tailed; FDR corrected), while no bins had significantly lower mean relative difference values using

TABLE 3 Number of white matter parcels retained in each dataset after removing the parcels that were not consistently detected across subjects in each dataset

	ABIDE-II	HCP	PPMI
Fiber clustering	1,274 (9,569 mm ³)	1,499 (10,088 mm ³)	1,373 (10,025 mm ³)
Cortical-parcellation-based	1968 (9,081 mm ³)	2,350 (8,853 mm ³)	2,269 (9,513 mm ³)

The parcels that were detected in at least 95% of the subjects in each dataset were retained (Section 2.4.1). The median of the mean volume of the retained parcels in each dataset is given in parentheses.

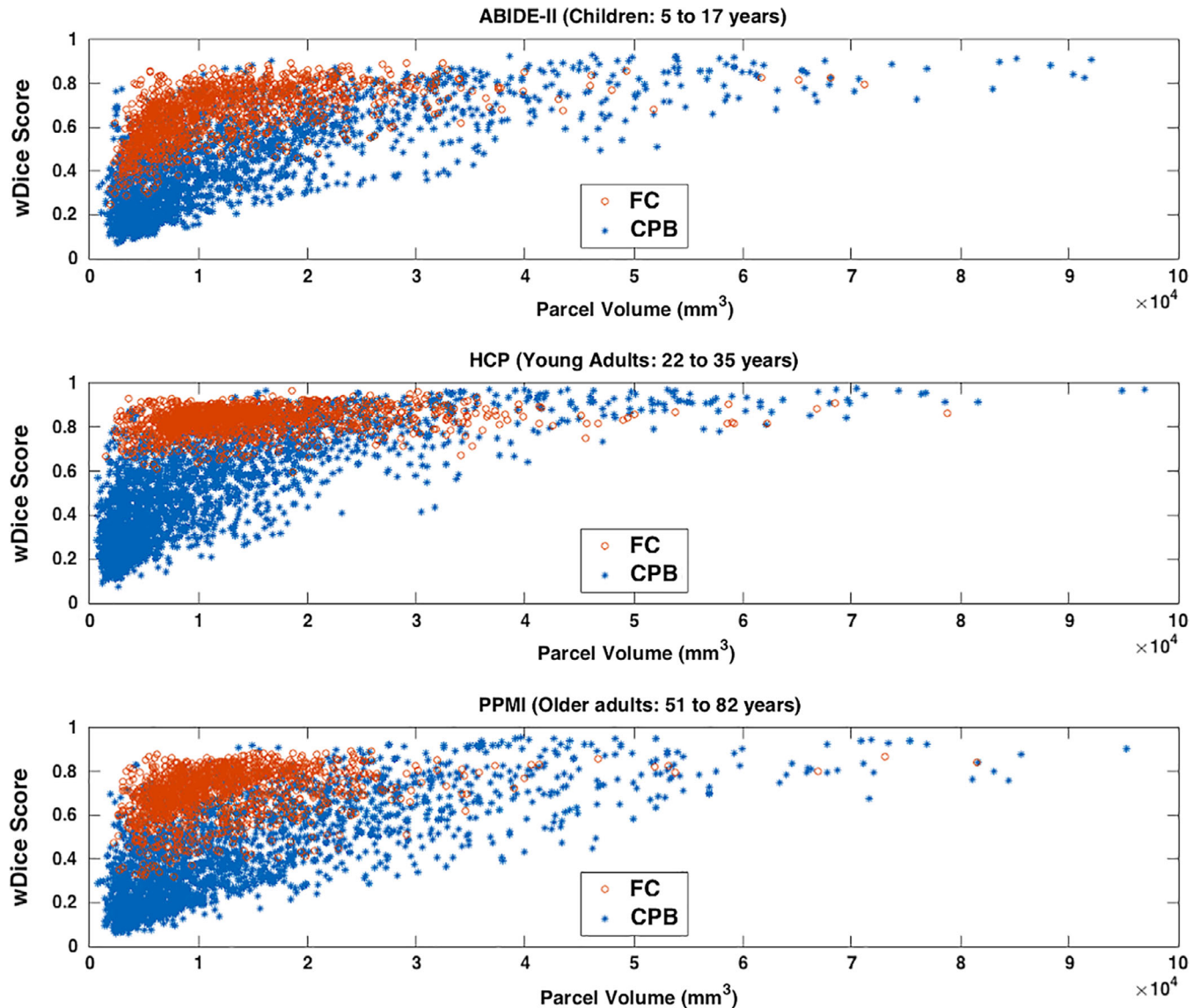


FIGURE 2 Volumetric overlap of white matter parcels computed from the test–retest dMRI data using the fiber clustering (FC) and cortical-parcellation-based (CPB) methods. Each plotted point represents one parcel and shows the parcel's mean wDice score versus the mean parcel volume across all subjects in one dataset. A high wDice score represents a high test–retest reproducibility [Color figure can be viewed at wileyonlinelibrary.com]

the cortical-parcellation-based method. (See Supporting Information S2 for the number of the retained bins in each dataset after removing the ones that had 0 parcels in either method.)

3.2 | Anatomical fiber tract parcellation results

3.2.1 | Volumetric overlap of anatomical fiber tracts

Figure 5 shows the mean wDice of the 45 anatomical fiber tracts in each dataset, after application of the fiber clustering and the cortical-parcellation-based methods. In the comparison using all 45 tracts, significantly higher mean wDice was found using the fiber clustering method than the cortical-parcellation-based method in all of the three datasets, with $p < 0.001$ (paired t -test, two-tailed). In the more fine-grained comparison using individual tracts, 32, 38, and 36 tracts, respectively, in the ABIDE-II, HCP and PPMI datasets, had significantly

higher wDice scores using the fiber clustering method, while 5, 6, and 6 tracts had significantly higher wDice scores using the cortical-parcellation-based method ($p < 0.05$, paired t -test, two-tailed; FDR corrected; Additional results on the tract volume are provided in Supporting Information S3.)

For visual quality assessment of the volumetric overlap performance, Figure 6 gives an visualization of the tracts obtained using the fiber clustering and cortical-parcellation-based methods by showing the ones with high or low volumetric overlaps. To create this visualization, we computed an average wDice score for each tract. This average was computed across the results from all methods and datasets in that tract (across the two methods and three datasets). This gave a total of 45 average wDice values, one per tract. We then identified the tracts with the maximum, the median and the minimum mean wDice scores, which were the corpus callosum 6 (CC6), the left

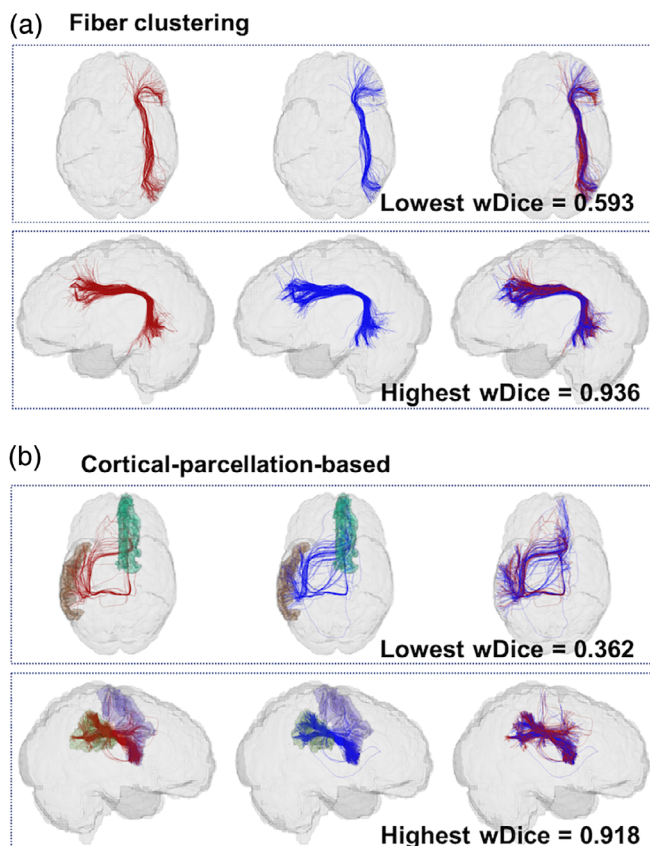


FIGURE 3 Visualization of volumetric overlap of white matter parcels (red—first scan; blue—second scan) obtained using the two parcellation methods. In the fiber clustering method, the parcel with the lowest wDice corresponds to the atlas cluster #728; the parcel with the highest wDice score corresponds to the atlas cluster #206. In the cortical-parcellation-based method, the parcel with the lowest wDice connects to the Freesurfer regions of the left middle temporal gyrus (brown) and right hemispheric superior frontal gyrus (dark green); the parcel with the highest wDice score connects to the Freesurfer regions of the right hemispheric precentral gyrus (purple) and the right hemispheric supramarginal gyrus (light green) [Color figure can be viewed at wileyonlinelibrary.com]

inferior occipito-frontal fasciculus (IOFF-L), the right striato-parietal (SP-R) tracts, respectively. In Figure 6, we display these three tracts from one HCP dataset.

3.2.2 | Relative difference of FA of anatomical fiber tracts

Figure 7 shows the mean relative difference of FA of the 45 anatomical fiber tracts in each dataset, after application of the fiber clustering and the WMQL methods. In the statistical comparison across all 45 tracts, two of the three datasets (ABIDE-II and PPMI) had significantly lower relative difference values in the fiber clustering method than in the cortical-parcellation-based method ($p = 0.032$ and 0.012 , respectively, paired t -test, two-tailed), while there was no significant difference between the two methods in the HCP dataset ($p = 0.620$, paired t -test, two-tailed). In the comparison on individual fiber tracts, 5, 1, and 7 tracts, respectively, in the ABIDE-II, HCP, and PPMI datasets, had significantly lower relative difference values using the fiber clustering method, while 1, 1 and 2 tracts had significantly lower

relative difference values using the cortical-parcellation-based method ($p < 0.05$, paired t -test, two-tailed; FDR corrected).

4 | DISCUSSION

In this work, we assessed test–retest reproducibility of two popular white matter tract parcellation strategies, including a white-matter-atlas-based fiber clustering method and a Freesurfer-based cortical-parcellation-based method. Overall, we found that the fiber clustering method had significantly higher reproducibility than the cortical-parcellation-based method in white matter parcellations for dividing the entire white matter into whole-brain parcels and identifying anatomical fiber tracts. When comparing all parcellated structures (either all whole-brain parcels or all anatomical tracts), the fiber clustering method obtained significantly higher reproducibility on both volumetric overlap and relative difference of FA in all of the three datasets. In more fine-grained comparisons on individual anatomical tracts, volumetric overlap and relative difference of FA were significantly higher using the fiber clustering method in 73.10% and 9.63% of tracts, respectively, on average across the three datasets. In contrast, volumetric overlap and relative difference of FA were significantly higher in 12.59% of tracts and 2.96% of tracts using the cortical-parcellation-based method. Below, we discuss several detailed observations regarding the comparison results for these two types of white matter parcellation.

We found that the test–retest reproducibility of white matter parcels obtained in whole brain white matter parcellation was highly related to their parcel volumes, such that parcels with larger volumes tended to be more reproducible. The cortical-parcellation-based method generated a white matter parcel based on its connected cortical or subcortical ROIs, where large-size ROIs likely generated parcels with large volumes. As a result, the reproducibility of a cortical-parcellation-based parcel was highly determined by the size of its connected ROIs, in agreement with other findings in the literature (Chamberland et al., 2017). For example, the most reproducible cortical-parcellation-based parcels connected to large-size Freesurfer regions such as the superior frontal gyrus and the rostral middle frontal gyrus (see Supporting Information S4). Similar results related to the ROI size in cortical-parcellation-based strategies have also been reported in several previous studies (Bonilha et al., 2015; Cheng et al., 2012). On the other hand, the fiber clustering method produced white matter parcels based on the white matter anatomy. It did not require any gray matter anatomical information; thus it would not be affected by the sizes of ROIs. For instance, we found that the fiber clustering parcels related to the frontal pole, which is a relatively small Freesurfer cortical ROI, were highly reproducible (see Supporting Information S4). However, while the cortical-parcellation-based method in general had the highest test–retest reproducibility on the largest volume parcels, the fiber clustering method did not produce parcels with volumes as large as the largest volume parcels from the cortical-parcellation-based method. For example, there were only a small number of fiber clustering parcels with volumes over $40,000 \text{ mm}^3$ (Figure 2).

We observed that the volumetric overlap of the anatomical fiber tracts was more sensitive than the diffusion FA measure for finding differences in test–retest reproducibility between the methods. There

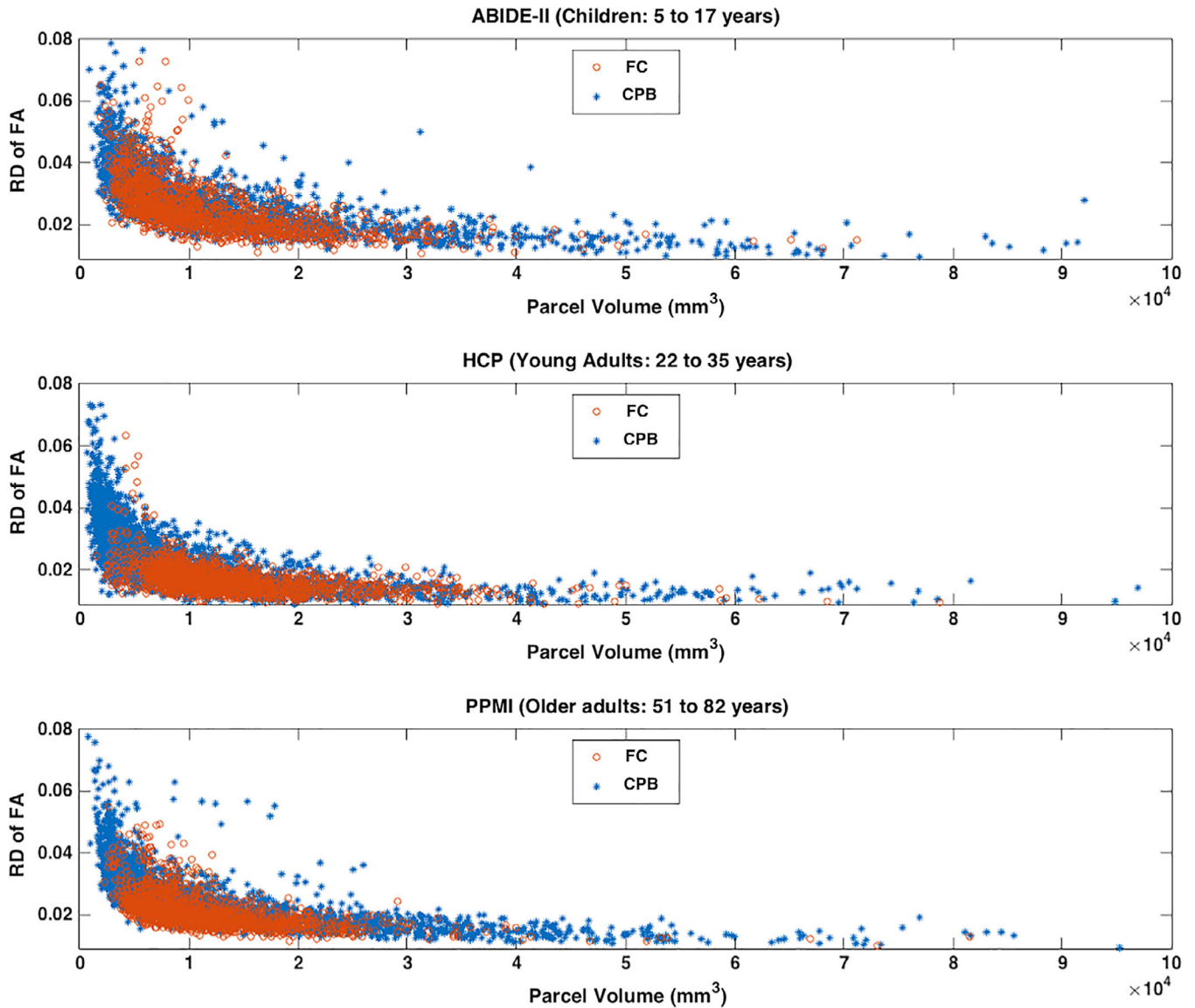


FIGURE 4 Reproducibility of diffusion FA measure of white matter parcels computed from the test–retest dMRI data using the fiber clustering (FC) and cortical-parcellation-based (CPB) methods. In the plots, each point represents the mean relative difference (RD) of FA versus the mean parcel volume across all subjects in one dataset. A low relative difference value represents a high test–retest reproducibility [Color figure can be viewed at wileyonlinelibrary.com]

were a higher number of significant differences between the two methods when using the wDice overlap score than relative difference of FA (Figure 5 vs. Figure 7). The wDice score was driven by non-overlapping voxels that were not intersected by the white matter structure in both test–retest scans, while the relative difference of FA was driven by the FA differences of these voxels. It is possible that the higher number of significant differences using wDice, when compared to relative difference of FA, can be attributed to the fact that the measured FA is averaged in the entire tract. wDice is expected to be more sensitive to small changes in the borders or edges of the fiber tract, which may not greatly affect the tract mean FA. Therefore, small differences in the voxels intersected by the fiber tract do not have a large effect on the measured mean FA. In the field of studying brain white matter properties (e.g., for analyzing disease analysis, understanding neurodevelopment, etc.), the reproducibility of diffusion

measures (such as FA and MD) is highly important. Our results suggested that the fiber clustering and cortical-parcellation-based methods performed comparably on a diffusion measure (Figure 7). For example, while there were a larger number of tracts with significantly lower relative difference values using the fiber clustering method in the ABIDE II (5 tracts versus 1 tract in the cortical-parcellation-based method) and PPMI (7 tracts versus 2 tracts in the cortical-parcellation-based method) datasets, both methods had the same number of tracts with significantly lower relative difference values in the HCP dataset (both had 1 tract).

In comparison to related work on test–retest reproducibility of white matter parcellation, we found that both of the fiber clustering and cortical-parcellation-based methods performed relatively well. While existing studies applied different evaluation criteria using different testing datasets (Besseling et al., 2012; Cheng et al., 2012;

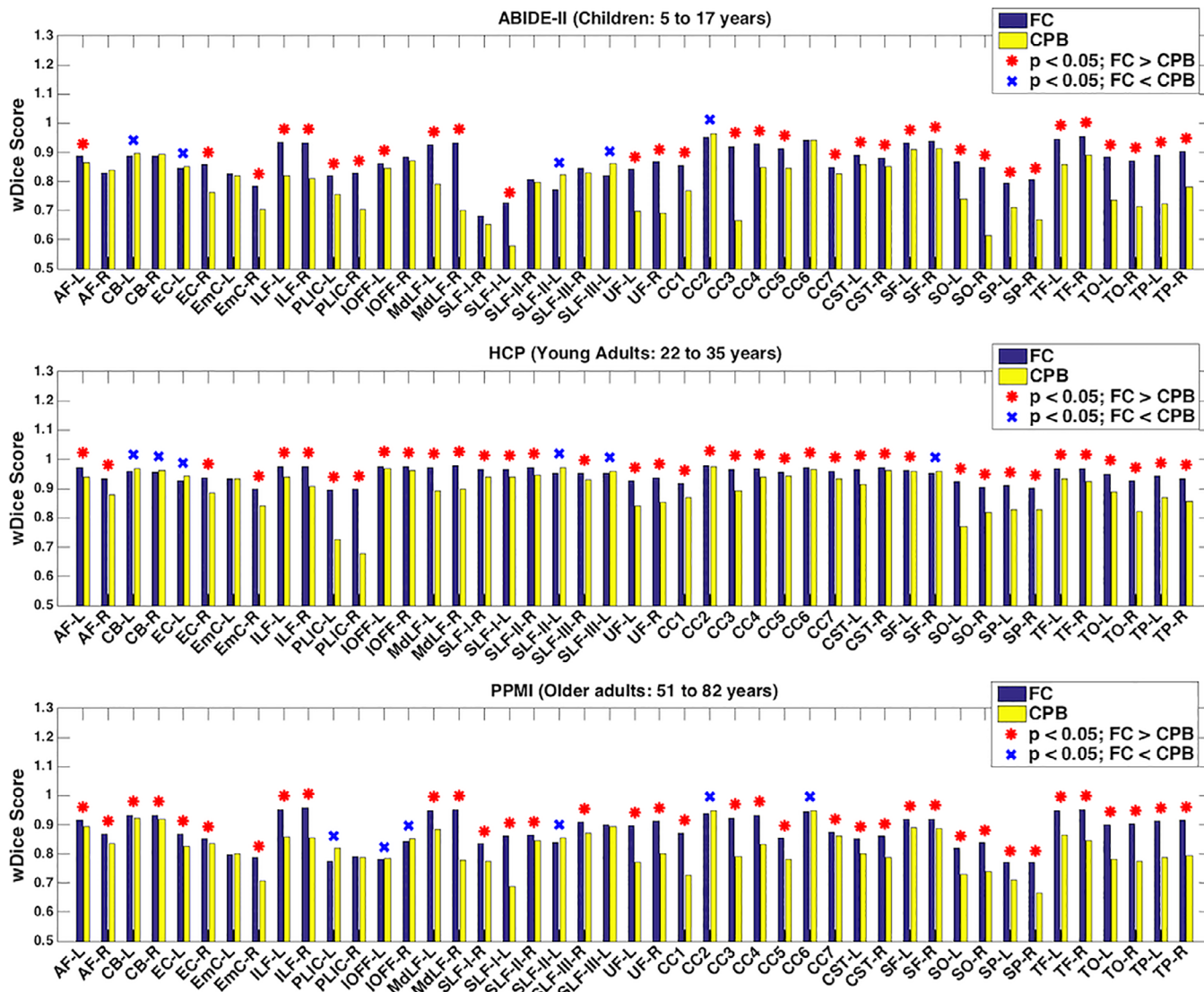


FIGURE 5 Volumetric overlap of anatomical fiber tracts identified from the test–retest dMRI data using the fiber clustering (FC) and cortical-parcellation-based (CPB) methods. Plots show the mean wDice score (averaged across subjects in each dataset) for each tract. A high wDice score represents a high test–retest reproducibility. The tracts with significantly higher mean wDice scores using the fiber clustering method are annotated with a red asterisk, while the tracts with significantly higher mean wDice scores using the cortical-parcellation-based method are annotated with a blue X [Color figure can be viewed at wileyonlinelibrary.com]

Ciccarelli et al., 2003; Cousineau et al., 2017; Duan et al., 2015; Kristo et al., 2013; Lin et al., 2013; Owen et al., 2015; Papinutto et al., 2013; Pfefferbaum et al., 2003; Smith et al., 2015; Vollmar et al., 2010; Wang et al., 2012; Yendiki et al., 2016; Zhao et al., 2015), one study performed volumetric-overlap-based experiments in a comparable way to the present study (Cousineau et al., 2017). In this work, Cousineau et al. studied test–retest reproducibility of a Freesurfer-based cortical-parcellation-based anatomical tract parcellation on PPMI data and suggested a threshold for a good wDice score to be 0.72 (based on an analysis of the mean wDice score across data in a healthy population; Cousineau et al., 2017). Although different parcellated tracts and different PPMI subject subsets were analyzed, 44 of 45 cortical-parcellation-based parcellated tracts and all fiber clustering parcellated tracts had mean wDice scores over 0.72 in our PPMI dataset, while only 8 of 28 parcellated tracts in (Cousineau et al., 2017) had mean wDice scores over this threshold. (This difference could potentially relate to the high

consistency of UKF tractography across different scan protocols and age groups (Zhang, Wu, Norton, et al., 2018). On average across all three datasets in our study, 87.41% of cortical-parcellation-based parcellated tracts and 99.26% of fiber clustering parcellated tracts had mean wDice scores over 0.72. In another study, Papinutto et al. measured test–retest reproducibility of FA of three tracts to evaluate across different acquisition dMRI parameters, and they reported that the lowest mean reproducibility error was around 2.4% (corresponding to a mean relative difference value of 0.012 in our study; Papinutto et al., 2013). In our study, on average across all the three datasets, the mean relative difference values were 0.0122 and 0.0139 using the fiber clustering and cortical-parcellation-based methods, respectively, which were close to the lowest value reported by Papinutto et al. In addition to comparison with existing work, we observed that changes in tract FA values between the test–retest scans ($(FA_{1st} - FA_{2nd})/FA_{1st}$) were relatively small (see Supporting Information S5 for details). On average

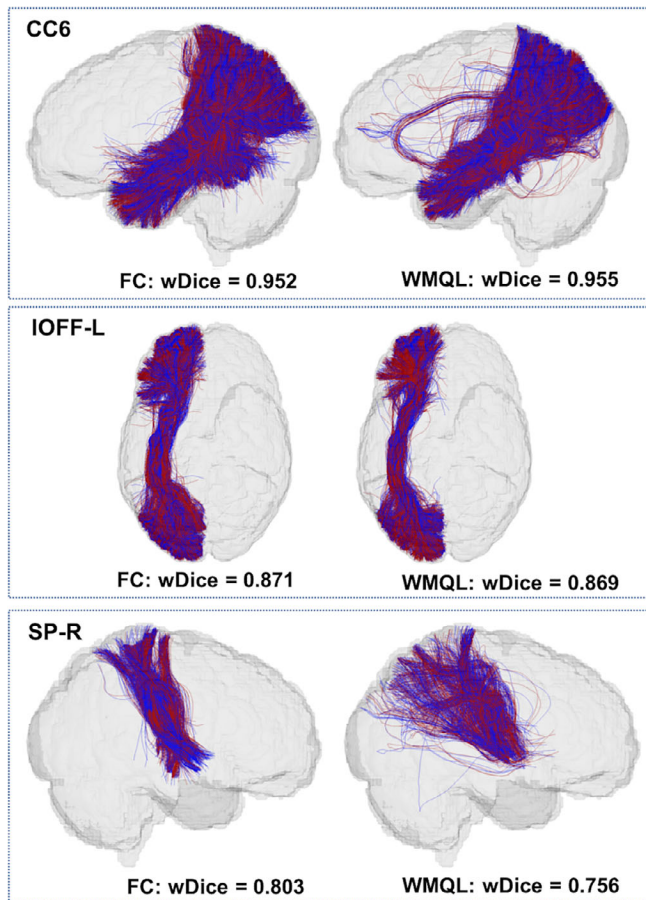


FIGURE 6 Visualization of volumetric overlap of example anatomical fiber tracts (red—first scan; blue—second scan) identified using the two parcellation methods. These three tracts have the maximum, the median and the minimum mean wDice scores. The tracts are corpus callosum 6 (CC6), left inferior occipito-frontal fasciculus (IOFF-L), and right striato-parietal (SP-R) tract [Color figure can be viewed at wileyonlinelibrary.com]

across the three datasets, FA changes of only 3.11% and 3.40% were observed in the fiber clustering and cortical-parcellation-based methods, respectively. These statistics showed that the cortical-parcellation-based and fiber clustering methods used in the present study obtained generally reproducible results.

Overall, the differences of the test–retest reproducibility results from the two white matter parcellation methods can be explained as follows. First, the methods had *different assumptions relative to the input tractography*. The cortical-parcellation-based method relied on particular points on the fibers, especially on fiber terminal regions. Parcellation results were thus sensitive to whether the fiber endpoints touched the cortical and subcortical ROI. This could be affected by multiple factors, for example, whether tractography could track through the low-anisotropy interface between gray matter and white matter or in deep gray matter regions near corticospinal fluid (CSF), where fiber endpoints are uncertain. The fiber clustering method, on the other hand, used all points on the fibers, that is, the full length of the fiber trajectory. In this way, fibers whose endpoints did not quite reach the cortex or the subcortical structures could nevertheless be parcellated. Using the entire fiber trajectory could also enable

localization of compact fiber clustering parcels, within which all fibers followed similar paths. This tendency toward compact fiber clustering parcels, versus the potentially more dispersed or spatially sparse parcels that were possible in the cortical-parcellation-based method (Figure 3), could also partially explain the higher volumetric overlap observed in the fiber clustering method. Second, the two methods applied *different image registration steps* to align the input tractography data to an atlas parcellation space. The cortical-parcellation-based method used multiple registration steps, including a registration between the subject-specific dMRI data and the subject-specific T1-weighted data, and a registration between the subject-specific T1-weighted data and the Freesurfer atlas. While there are sophisticated tools to compute these registrations (Avants et al., 2009; Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012), the performance is limited by nontrivial factors. For example, the intermodality registration between dMRI and T1-weighted can be affected by differences in image resolutions (Malinsky et al., 2013) echo-planar imaging (EPI) distortion in dMRI data (Albi et al., 2018). Large individual anatomical variations with respect to the atlas population could also affect, or even cause to fail, the subject-specific T1 registration to the atlas. For example, Freesurfer has been shown to have limited success in neonates (Makropoulos et al., 2018) and patients with brain tumors or other structural lesions (Zhang et al., 2017). In contrast, the fiber clustering method needed only one intra-modality registration step between the subject-specific tractography data and the atlas tractography data. This registration, in combination with robust fiber spectral embedding in the atlas space, has been demonstrated to enable successful white matter parcellation across subjects from different populations, including neonates, young children, and brain tumor patients, who had large neuroanatomical variations to the atlas population (Zhang, Wu, Norton, et al., 2018). Third, the two methods adopted *different ways to handle false positive fibers* that have been suggested to be a contributing factor affecting white matter parcellation reproducibility (Maier-Hein et al., 2017). In the present study, we applied a multi-fiber UKF tractography method to increase the sensitivity in tracking crossing fibers (Baumgartner et al., 2012; Chen et al., 2015, 2016; Liao et al., 2017). The high sensitivity has been suggested to be important to reduce false negatives, but at the expense of increased false positives (Maier-Hein et al., 2017; Thomas et al., 2014). Therefore, the UKF method, as well as other sensitive fiber tracking methods (Christiaens et al., 2015; Jeurissen, Tournier, Dhollander, Connelly, & Sijbers, 2014), may introduce more false positive or anatomically incorrect errors compared to a standard single-fiber diffusion tensor fiber tracking method. Prior anatomical knowledge is often used to exclude false positive fibers by employing additional constraints and expert judgment (Conturo et al., 1999; Huang, Zhang, van Zijl, & Mori, 2004; Yeh et al., 2018). In the cortical-parcellation-based method, the WMQL tract definitions, which were originally designed for standard single-tensor tractography (Demian Wassermann et al., 2016), were improved following query testing and modification (by NM and colleagues) for the more sensitive UKF by including additional ROIs to constrain fiber selection (Sydnor et al., 2018). In the FC method, false positive fibers in the atlas were annotated and rejected via expert judgment to ameliorate potential subject-specific false positive fibers that were inconsistent with respect to

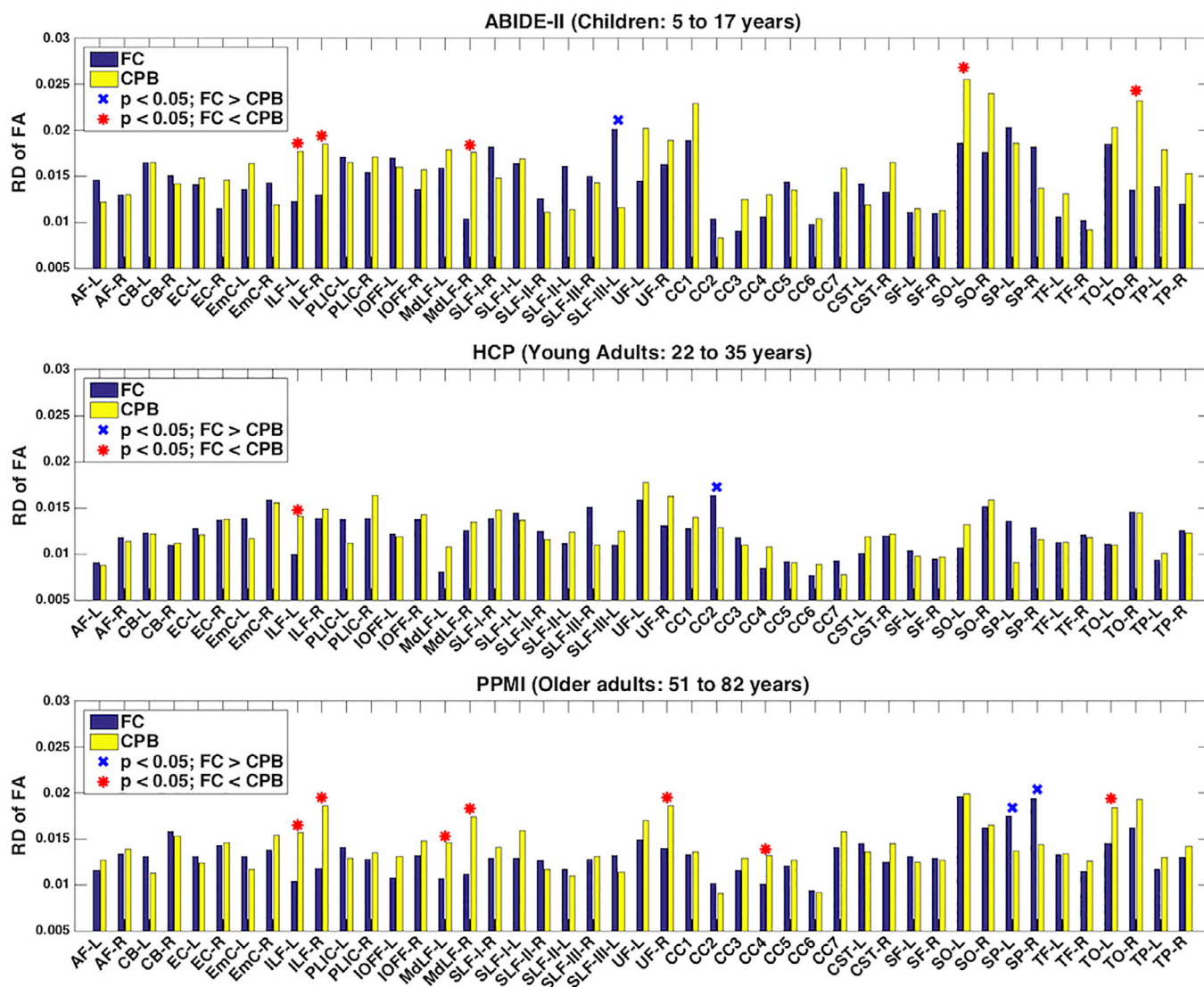


FIGURE 7 Reproducibility of diffusion FA measure of anatomical fiber tracts identified from the test-retest dMRI data using the fiber clustering (FC) and cortical-parcellation-based (CPB) methods. A low relative difference (RD) value represents a high test-retest reproducibility. The tracts with significantly lower mean relative difference scores using the fiber clustering method are annotated with a red asterisk, while the tracts with significantly lower mean relative difference scores using the cortical-parcellation-based method are annotated with a blue X [Color figure can be viewed at wileyonlinelibrary.com]

known neuroanatomical knowledge (Zhang, Wu, Norton, et al., 2018). In addition to the prior anatomical knowledge annotated in the atlas, the fiber clustering method also included a data-driven false positive fiber removal for rejection of fibers that have improbable fiber geometric trajectory (O'Donnell et al., 2017; Zhang, Wu, Norton, et al., 2018), similar to the processing applied in many other fiber clustering methods (Côté, Garyfallidis, Laroche, & Descoteaux, 2015; Xia, Turken, Whitfield-Gabrieli, & Gabrieli, 2005; Ziyang et al., 2009). This removal of false positive fibers can be one factor that improves the reproducibility of the wDice score in the FC method.

We also found that the test-retest reproducibility from the three datasets was different. In general, the HCP dataset had higher performance than the other two datasets. This likely relates to the much higher quality of the HCP acquisition (Table 1). It is well known that estimation of diffusion MRI parameters is more robust with a higher number of gradient directions and higher signal-to-noise ratio (Farrell et al., 2007; Jones, 2004). The data quality may also explain that the

HCP dataset had a larger number of retained white matter parcels than the other two datasets in both the fiber clustering and cortical-parcellation-based methods (Table 2).

Test-retest reproducibility is considered to be a good indicator of the reliability of white matter parcellation for potential clinical applications, as well as the study of large datasets for neuroscientific research. Having highly reproducible parcellation is a prerequisite for clinical applications such as neurosurgery and neurology (Kristo et al., 2013). Test-retest reproducibility is also important to determine the sensitivity of tractography to reveal pathological abnormalities and changes over time (Besseling et al., 2012). Multiple studies have used test-retest reproducibility to assess predictive power in studying brain longitudinal changes (Jovicich et al., 2014; Keihaninejad et al., 2013; Lin et al., 2013). In our study, we showed a highly reproducible white matter parcellation using the fiber clustering method; thus, we suggested this method could be reproducibly applied for clinical applications and large dataset analysis. These applications can include

white matter parcellation for neurosurgical planning (O'Donnell et al., 2017), parcellation of scans from different time points to track white matter changes, and analysis of very large diffusion MRI datasets that will soon become available (Alexander et al., 2017; Casey et al., 2018; Thompson et al., 2017).

Potential future directions and limitations of the current work are as follows. First, the two compared methods produced different numbers of parcels, with different distributions of parcel volumes (though the median parcel volumes were not significantly different). The scale of the white matter parcellation (i.e., the number of parcels, or the volume of those parcels) is an important factor in the success of the parcellation for a particular application. Many studies have shown that different parcellation scales can provide a better description of local brain regions (Cammoun et al., 2012; Hagmann et al., 2007; Liu et al., 2017; Zhang, Savadjiev, Cai, et al., 2018). The scale of a white matter parcellation also affects the intersubject parcellation consistency (Zhang, Norton, et al., 2017). Future work could include an investigation of different white matter parcellation scales on test-retest reproducibility. Second, the aim of this study was to compare test-retest reproducibility of different white matter parcellation methods on several independently acquired datasets; thus we did not perform any statistical analyses across the different populations. There have been studies that investigate differences of reproducibility measures, for example, between disease versus healthy populations (Cousineau et al., 2017; Lin et al., 2013) and between data with different acquisitions (Papinutto et al., 2013), using cortical-parcellation-based strategies. Future work could include applying fiber clustering strategies for such statistical analyses. Third, in the present study, we chose volumetric overlap and FA to investigate white matter parcellation test-retest reproducibility. While these two measures are relatively representative test-retest reproducibility measures and have been used in many related studies (Cousineau et al., 2017; Besseling et al., 2012; Ciccarelli et al., 2003; Kristo et al., 2013; Papinutto et al., 2013; Pfefferbaum et al., 2003; Vollmar et al., 2010), there are many other measures available, for example, MD, apparent fiber density and number of fiber orientations (Cousineau et al., 2017; Kristo et al., 2013; Kuhn et al., 2016; Papinutto et al., 2013). In Supporting Information S1, we provide additional experimental results from relative difference of MD on whole brain white matter parcellation and anatomical fiber tract parcellation. The results in general agree with our overall finding, that is, the fiber clustering method generates significantly more reproducible white matter parcellations than the cortical-parcellation-based method. However, we noticed that the performance on MD was slightly different from that on FA. For example, there are more significantly different tracts between the two methods using MD (11, 3, and 20, respectively, in the three datasets) compared to FA (6, 2, and 9, respectively, in the three datasets). This could potentially be related to different levels of fiber specificities of FA and MD. Therefore, a further investigation could be done by including more test-retest reproducibility measures and a comparison between them. Fourth, we performed comparison of the fiber clustering and cortical-parcellation-based methods using a deterministic UKF tractography method, which is highly consistent in tracking fibers in dMRI data from independently acquired populations across ages, health conditions and image acquisitions (Zhang, Wu, Norton, et al., 2018). However,

many other tractography methods (such as probabilistic (Jeurissen, Leemans, Jones, Tournier, & Sijbers, 2011), global (Christiaens et al., 2015), and multi-tissue (Jeurissen et al., 2014) fiber tracking methods) potentially could generate improved white matter parcellation test-retest reproducibility. Fifth, in the present study, we compared two fiber clustering and cortical-parcellation-based strategies that are widely used but relatively traditional approaches. In the past few years, there have been methods designed to improve test-retest reproducibility for constructing cortical-parcellation-based connectomes by including additional pre- and/or postprocessing steps, for example, dilating gray matter regions (Zhang, Descoteaux, Zhang, et al., 2018), constructing continuous connectome matrices (Moyer, Gutman, Faskowitz, Jahanshad, & Thompson, 2017), and filtering out implausible fiber streamlines (Smith et al., 2015). A further study could include comparison of test-retest reproducibility between the fiber clustering and cortical-parcellation-based strategies with advanced processing.

5 | CONCLUSION

Our experimental results in general indicate that the fiber clustering method generates significantly more reproducible white matter parcellations than the cortical-parcellation-based method. However, both methods have high performance when compared to existing studies of reproducibility of fiber tract anatomical parcellation.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding provided by the following National Institutes of Health (NIH) grants: P41 EB015902, P41 EB015898, R01 MH074794, R01 MH097979, U01 CA199459, and R03 NS088301.

ORCID

Fan Zhang  <https://orcid.org/0000-0002-5032-6039>

Yogesh Rathi  <https://orcid.org/0000-0002-9946-2314>

Lauren J. O'Donnell  <https://orcid.org/0000-0003-0197-7801>

REFERENCES

- Albi, A., Meola, A., Zhang, F., Kahali, P., Rigolo, L., Tax, C. M. W., et al. (2018). Image registration to compensate for EPI distortion in patients with brain tumors: An evaluation of tract-specific effects. *Journal of Neuroimaging: Official Journal of the American Society of Neuroimaging*, 28(2), 173–182.
- Alexander, A. L., Lee, J. E., Lazar, M., Boudos, R., DuBray, M. B., Oakes, T. R., et al. (2007). Diffusion tensor imaging of the corpus callosum in autism. *NeuroImage*, 34(1), 61–73.
- Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., ... Milham, M. P. (2017). An open resource for Transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4 (December), 170181.
- Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B., & Zilles, K. (1999). Broca's region revisited: Cytoarchitecture and intersubject variability. *The Journal of Comparative Neurology*, 412(2), 319–341.
- Avants, B. B., Tustison, N., & Song, G. (2009). Advanced normalization tools (ANTS). *The Insight Journal*, 2, 1–35.

- Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, 66(1), 259–267.
- Basser, P. J., Pajevic, S., Pierpaoli, C., Duda, J., & Aldroubi, A. (2000). In vivo fiber Tractography using DT-MRI data. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine*, 44(4), 625–632.
- Bassett, D. S., & Bullmore, E. T. (2016). Small-world brain networks revisited. *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 23, 499–516. <https://doi.org/10.1177/1073858416667720>
- Bastiani, M., Shah, N. J., Goebel, R., & Roebroeck, A. (2012). Human cortical connectome reconstruction from diffusion weighted MRI: The effect of Tractography algorithm. *NeuroImage*, 62(3), 1732–1749.
- Baumgartner, Christian, O. Michailovich, J. Levitt, O. Pasternak, S. Bouix, C. Westin, and Yogesh Rathi. 2012. A unified Tractography framework for comparing diffusion models on clinical scans. In *Computational diffusion MRI workshop of MICCAI*, Nice, pp. 27–32.
- 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(1), 289–300.
- Besseling, R. M. H., Jansen, J. F. A., Overvliet, G. M., Vaessen, M. J., Braakman, H. M. H., Hofman, P. A. M., ... Backes, W. H. (2012). Tract specific reproducibility of Tractography based morphology and diffusion metrics. *PLoS One*, 7(4), e34125.
- Besson, P., Lopes, R., Leclerc, X., Derambure, P., & Tyvaert, L. (2014). Intra-subject reliability of the high-resolution whole-brain structural connectome. *NeuroImage*, 102(2), 283–293.
- Boniilha, L., Gleichgerrcht, E., Fridriksson, J., Rorden, C., Breedlove, J. L., Nesland, T., ... Focke, N. K. (2015). Reproducibility of the structural brain connectome derived from diffusion tensor imaging. *PLoS One*, 10(8), e0135247.
- Buchanan, C. R., Pernet, C. R., Gorgolewski, K. J., Storkey, A. J., & Bastin, M. E. (2014). Test–retest reliability of structural brain networks from diffusion MRI. *NeuroImage*, 86(February), 231–243.
- Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, 10(3), 186–198.
- Cammoun, L., Gigandet, X., Meskaldji, D., Thiran, J. P., Sporns, O., Do, K. Q., ... Hagmann, P. (2012). Mapping the human connectome at multiple scales with diffusion Spectrum MRI. *Journal of Neuroscience Methods*, 203(2), 386–397.
- Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., et al. (2018). The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32, 43–54.
- Chamberland, M., Girard, G., Bernier, M., Fortin, D., Descoteaux, M., & Whittingstall, K. (2017). On the origin of individual functional connectivity variability: The role of White matter architecture. *Brain Connectivity*, 7(8), 491–503.
- Cheng, H., Yang, W., Sheng, J., Kronenberger, W. G., Mathews, V. P., Hummer, T. A., & Saykin, A. J. (2012). Characteristics and variability of structural networks derived from diffusion tensor imaging. *NeuroImage*, 61(4), 1153–1164.
- Chen, Z., Tie, Y., Olubiyi, O., Rigolo, L., Mehrtash, A., Norton, I., ... O'Donnell, L. J. (2015). Reconstruction of the arcuate fasciculus for surgical planning in the setting of Peritumoral edema using two-tensor unscented Kalman filter Tractography. *NeuroImage: Clinical*, 7(March), 815–822.
- Chen, Z., Tie, Y., Olubiyi, O., Zhang, F., Mehrtash, A., Rigolo, L., ... O'Donnell, L. J. (2016). Corticospinal tract modeling for neurosurgical planning by tracking through regions of Peritumoral edema and crossing fibers using two-tensor unscented Kalman filter Tractography. *International Journal of Computer Assisted Radiology and Surgery*, 11(8), 1475–1486.
- Christiaens, D., Reisert, M., Dhollander, T., Sunaert, S., Suetens, P., & Maes, F. (2015). Global Tractography of multi-shell diffusion-weighted imaging data using a multi-tissue model. *NeuroImage*, 123(December), 89–101.
- Ciccarelli, O., Catani, M., Johansen-Berg, H., Clark, C., & Thompson, A. (2008). Diffusion-based Tractography in neurological disorders: Concepts, applications, and future developments. *Lancet Neurology*, 7(8), 715–727.
- Ciccarelli, O., Parker, G. J. M., Toosy, A. T., Wheeler-Kingshott, C. A. M., Barker, G. J., Boulby, P. A., ... Thompson, A. J. (2003). From diffusion Tractography to quantitative White matter tract measures: A reproducibility study. *NeuroImage*, 18(2), 348–359.
- Conturo, T. E., Lori, N. F., Cull, T. S., Akbudak, E., Snyder, A. Z., Shimony, J. S., ... Raichle, M. E. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 96(18), 10422–10427.
- Côté, Marc-Alexandre, Eleftherios Garyfallidis, Hugo Larochelle, and Maxime Descoteaux. 2015. Cleaning up the mess: Tractography outlier removal using hierarchical QuickBundles clustering. *Proceedings of the International Society for Magnetic Resonance in Medicine ... Scientific Meeting and Exhibition*. International Society for Magnetic Resonance in Medicine. Scientific Meeting and Exhibition. <http://scil.dinf.usherbrooke.ca/wp-content/papers/cote-et-al-ismrm15.pdf>.
- Cousineau, M., Jodoin, P.-M., Morency, F. C., Rozanski, V., Grand'Maison, M., Bedell, B. J., & Descoteaux, M. (2017). A test-retest study on Parkinson's PPMI dataset yields statistically significant White matter fascicles. *NeuroImage: Clinical*, 16, 222–233.
- Dennis, Emily L., Neda Jahanshad, Arthur W. Toga, Katie L. McMahon, Greig I. de Zubicaray, Nicholas G. Martin, Margaret J. Wright, and Paul M. Thompson. 2012. Test-retest reliability of graph theory measures of structural brain connectivity. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 15 (Pt 3), pp. 305–312).
- Santis, D., Silvia, M. D., Bells, S., Assaf, Y., & Jones, D. K. (2014). Why diffusion tensor MRI does well only some of the time: Variance and covariance of White matter tissue microstructure attributes in the living human brain. *NeuroImage*, 89, 35–44.
- Descoteaux, M., Angelino, E., Fitzgibbons, S., & Deriche, R. (2007). Regularized, fast, and robust analytical Q-ball imaging. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine*, 58(3), 497–510.
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into Gyral based regions of interest. *NeuroImage*, 31(3), 968–980.
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.
- Martino, D., Adriana, D. O'. C., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., et al. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, 4, 170010.
- Ding, Z., Gore, J. C., & Anderson, A. W. (2003). Classification and quantification of neuronal fiber pathways using diffusion tensor MRI. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine*, 49(4), 716–721.
- Duan, F., Zhao, T., He, Y., & Shu, N. (2015). Test-retest reliability of diffusion measures in cerebral White matter: A multiband diffusion MRI study. *Journal of Magnetic Resonance Imaging: JMRI*, 42(4), 1106–1116.
- Duda, J. T., Cook, P. A., & Gee, J. C. (2014). Reproducibility of graph metrics of human brain structural networks. *Frontiers in Neuroinformatics*, 8, 46.
- Essayed, W. I., Zhang, F., Prashin, U., Rees Cosgrove, G., Golby, A. J., & O'Donnell, L. J. (2017). White matter Tractography for neurosurgical planning: A topography-based review of the current state of the art. *NeuroImage: Clinical*, 15(June), 659–672.
- Farrell, J. A. D., Landman, B. A., Jones, C. K., Smith, S. A., Prince, J. L., van Zijl, P. C. M., & Mori, S. (2007). Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5T. *Journal of Magnetic Resonance Imaging: JMRI*, 26(3), 756–767.
- Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781.
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3), 341–355.
- Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, 8(4), 272–284.
- Garyfallidis, E., Brett, M., Correia, M. M., Williams, G. B., & Nimmo-Smith, I. (2012). QuickBundles, a method for Tractography simplification. *Frontiers in Neuroscience*, 6, 175.
- Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., ... Descoteaux, M. (2018). Recognition of White matter bundles using

- local and global streamline-based registration and clustering. *NeuroImage*, 170, 283–295.
- Ge, Bao, Lei Guo, Tuo Zhang, Dajiang Zhu, Kaiming Li, Xintao Hu, Junwei Han, and Tianming Liu. 2012. Group-wise consistent fiber clustering based on multimodal connectonal and functional profiles. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 15 (Pt 3), pp. 485–492).
- Glasser, M. F., Sotiropoulos, S. N., Anthony Wilson, J., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80, 105–124.
- Golby, A. J., Kindlmann, G., Norton, I., Yarmarkovich, A., Pieper, S., & Kikinis, R. (2011). Interactive diffusion tensor Tractography visualization for neurosurgical planning. *Neurosurgery*, 68(2), 496–505.
- Gong, G., He, Y., Concha, L., Lebel, C., Gross, D. W., Evans, A. C., & Beaulieu, C. (2009). Mapping anatomical connectivity patterns of human cerebral cortex using in vivo diffusion tensor imaging Tractography. *Cerebral Cortex*, 19(3), 524–536.
- Gong, S., Zhang, F., Norton, I., Essayed, W. I., Unadkat, P., Rigolo, L., ... O'Donnell, L. J. (2018). Free water modeling of Peritumoral edema using multi-fiber Tractography: Application to tracking the arcuate fasciculus for neurosurgical planning. *PLoS One*, 13(5), e0197056.
- Goodlett, Casey, Brad Davis, Remi Jean, John Gilmore, and Guido Gerig. 2006. Improved correspondence for DTI population studies via unbiased atlas building. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol 9, pp. 260–267).
- Guevara, P., Duclap, D., Poupon, C., Marrakchi-Kacem, L., Fillard, P., Le Bihan, D., ... Mangin, J.-F. (2012). Automatic fiber bundle segmentation in massive Tractography datasets using a multi-subject bundle atlas. *NeuroImage*, 61(4), 1083–1099.
- Hagmann, P., Kurant, M., Gigandet, X., Thiran, P., Wedeen, V. J., Meuli, R., & Thiran, J.-P. (2007). Mapping human whole-brain structural networks with diffusion MRI. *PLoS One*, 2(7), e597.
- Heiervang, E., Behrens, T. E. J., Mackay, C. E., Robson, M. D., & Johansen-Berg, H. (2006). Between session reproducibility and between subject variability of diffusion MR and Tractography measures. *NeuroImage*, 33(3), 867–877.
- Huang, H., Zhang, J., van Zijl, P. C. M., & Mori, S. (2004). Analysis of noise effects on DTI-based Tractography using the brute-force and multi-ROI approach. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine*, 52(3), 559–565.
- Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., ... Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, 111(2), 823–828.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, 62(2), 782–790.
- Jeurissen, B., Leemans, A., Jones, D. K., Tournier, J.-D., & Sijbers, J. (2011). Probabilistic fiber tracking using the residual bootstrap with constrained spherical deconvolution. *Human Brain Mapping*, 32(3), 461–479.
- Jeurissen, B., Tournier, J.-D., Dhollander, T., Connelly, A., & Sijbers, J. (2014). Multi-tissue constrained spherical deconvolution for improved analysis of multi-Shell diffusion MRI data. *NeuroImage*, 103(December), 411–426.
- Jin, Y., Shi, Y., Liang, Z., Gutman, B. A., de Zubicaray, G. I., McMahon, K. L., ... Thompson, P. M. (2014). Automatic clustering of White matter fibers in brain diffusion MRI with an application to genetics. *NeuroImage*, 100, 75–90.
- Jones, D. K. (2004). The effect of gradient sampling schemes on measures derived from diffusion tensor MRI: A Monte Carlo study. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine/Society of Magnetic Resonance in Medicine*, 51(4), 807–815.
- Jovicich, J., Marizzoni, M., Bosch, B., Bartrés-Faz, D., Arnold, J., Benninghoff, J., ... Frisoni, G. B. (2014). Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects. *NeuroImage*, 101, 390–403.
- Keihaninejad, S., Zhang, H., Ryan, N. S., Malone, I. B., Marc, M., Jorge Cardoso, M., ... Ourselin, S. (2013). An unbiased longitudinal analysis framework for tracking White matter changes using diffusion tensor imaging with application to Alzheimer's disease. *NeuroImage*, 72, 153–163.
- Kristo, G., Leemans, A., de Gelder, B., Raemaekers, M., Rutten, G.-J., & Ramsey, N. (2013). Reliability of the corticospinal tract and arcuate fasciculus reconstructed with DTI-based Tractography: Implications for clinical practice. *European Radiology*, 23(1), 28–36.
- Kuhn, T., Gullett, J. M., Nguyen, P., Boutzoukas, A. E., Ford, A., Colon-Perez, L. M., ... Bauer, R. M. (2016). Test-retest reliability of high angular resolution diffusion imaging acquisition within medial temporal lobe connections assessed via tract based spatial statistics, probabilistic Tractography and a novel graph theory metric. *Brain Imaging and Behavior*, 10(2), 533–547.
- Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O., & Toews, M. (2017). Fiberprint: A subject fingerprint based on sparse code pooling for White matter fiber analysis. *NeuroImage*, 158, 242–259.
- Lefranc, S., Roca, P., Perrot, M., Poupon, C., Le Bihan, D., Mangin, J.-F., & Rivière, D. (2016). Groupwise connectivity-based Parcellation of the whole human cortical surface using watershed-driven dimension reduction. *Medical Image Analysis*, 30, 11–29.
- Liao, R., Ning, L., Chen, Z., Rigolo, L., Gong, S., Pasternak, O., ... O'Donnell, L. J. (2017). Performance of unscented Kalman filter Tractography in edema: Analysis of the two-tensor model. *NeuroImage: Clinical*, 15, 819–831.
- Lin, C.-C., Tsai, M.-Y., Lo, Y.-C., Liu, Y.-J., Tsai, P.-P., Wu, C.-Y., ... Chung, H.-W. (2013). Reproducibility of corticospinal diffusion tensor Tractography in Normal subjects and Hemiparetic stroke patients. *European Journal of Radiology*, 82(10), e610–e616.
- Liu, X., Lauer, K. K., Douglas Ward, B., Roberts, C. J., Liu, S., Gollapudy, S., et al. (2017). Fine-grained Parcellation of brain connectivity improves differentiation of states of consciousness during graded Propofol sedation. *Brain Connectivity*, 7(6), 373–381.
- Madden, D. J., Whiting, W. L., Huettel, S. A., White, L. E., MacFall, J. R., & Provenzale, J. M. (2004). Diffusion tensor imaging of adult age differences in cerebral White matter: Relation to response time. *NeuroImage*, 21(3), 1174–1181.
- Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., ... Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion Tractography. *Nature Communications*, 8(1), 1349.
- Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., ... Lenz, G. (2018). The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, 173, 88–112.
- Malcolm, J. G., Shenton, M. E., & Rathi, Y. (2010). Filtered multitensor Tractography. *IEEE Transactions on Medical Imaging*, 29(9), 1664–1675.
- Malinsky, M., Peter, R., Hodneland, E., Lundervold, A. J., Lundervold, A., & Jan, J. (2013). Registration of FA and T1-weighted MRI data of healthy human brain based on template matching and normalized cross-correlation. *Journal of Digital Imaging*, 26(4), 774–785.
- Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., ... Taylor, P. (2011). The Parkinson progression marker initiative (PPMI). *Progress in Neurobiology*, 95(4), 629–635.
- Moyer, D., Gutman, B. A., Faskowitz, J., Jahanshad, N., & Thompson, P. M. (2017). Continuous representations of brain connectivity using spatial point processes. *Medical Image Analysis*, 41, 32–39.
- Nimsky, C., Ganslandt, O., Dorit, M., Gregory Sorensen, A., & Fahlbusch, R. (2006). Intraoperative visualization of the pyramidal tract by diffusion-tensor-imaging-based fiber tracking. *NeuroImage*, 30(4), 1219–1229.
- Ning, L., Laun, F., Gur, Y., DiBella, E. V. R., Deslauriers-Gauthier, S., Megherbi, T., et al. (2015). Sparse reconstruction challenge for diffusion MRI: Validation on a physical phantom to determine which acquisition scheme and analysis method to use? *Medical Image Analysis*, 26(1), 316–331.
- Ning, L., Setsompop, K., Michailovich, O., Makris, N., Shenton, M. E., Westin, C.-F., & Rathi, Y. (2016). A joint compressed-sensing and super-resolution approach for very high-resolution diffusion imaging. *NeuroImage*, 125(January), 386–400.
- O'Donnell, L. J., Golby, A. J., & Westin, C.-F. (2013). Fiber clustering versus the Parcellation-based connectome. *NeuroImage*, 80(October), 283–289.
- O'Donnell, L. J., & Pasternak, O. (2015). Does diffusion MRI tell us anything about the White matter? An overview of methods and pitfalls. *Schizophrenia Research*, 161(1), 133–141.

- O'Donnell, L. J., Suter, Y., Rigolo, L., Kahali, P., Zhang, F., Norton, I., et al. (2017). Automated White matter fiber tract identification in patients with brain tumors. *NeuroImage: Clinical*, 13, 138–153.
- O'Donnell, Lauren J., William M., Wells, Alexandra J., Golby, and Carl-Fredrik, Westin. 2012. Unbiased groupwise registration of white matter tractography. In N. Ayache, H. Delingette, P. Golland, & K. Mori (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2012* (pp. 123–130). Lecture Notes in Computer Science. Vol. 7512. Berlin Heidelberg: Springer.
- O'Donnell, L. J., & Westin, C.-F. (2007). Automatic Tractography segmentation using a high-dimensional White matter atlas. *IEEE Transactions on Medical Imaging*, 26(11), 1562–1575.
- Owen, J. P., Chang, Y. S., & Mukherjee, P. (2015). Edge density imaging: Mapping the anatomic embedding of the structural connectome within the White matter of the human brain. *NeuroImage*, 109(April), 402–417.
- Pannek, K., Scheck, S. M., Colditz, P. B., Boyd, R. N., & Rose, S. E. (2014). Magnetic resonance diffusion Tractography of the preterm infant brain: A systematic review. *Developmental Medicine and Child Neurology*, 56(2), 113–124.
- Papinutto, N. D., Maule, F., & Jovicich, J. (2013). Reproducibility and biases in high field brain diffusion MRI: An evaluation of acquisition and analysis variables. *Magnetic Resonance Imaging*, 31(6), 827–839.
- Pfefferbaum, A., Adalsteinsson, E., & Sullivan, E. V. (2003). Replicability of diffusion tensor imaging measurements of fractional anisotropy and trace in brain. *Journal of Magnetic Resonance Imaging: JMIRI*, 18(4), 427–433.
- Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A., & Di Chiro, G. (1996). Diffusion tensor MR imaging of the human brain. *Radiology*, 201(3), 637–648.
- Piper, R. J., Yoong, M. M., Kandasamy, J., & Chin, R. F. (2014). Application of diffusion tensor imaging and Tractography of the optic radiation in anterior temporal lobe resection for epilepsy: A systematic review. *Clinical Neurology and Neurosurgery*, 124, 59–65.
- Prasad, G., Joshi, S. H., Jahanshad, N., Villalón-Reina, J., Aganj, I., Lenglet, C., ... Thompson, P. M. (2014). Automatic clustering and population analysis of White matter tracts using maximum density paths. *NeuroImage*, 97(August), 284–295.
- Reddy, C. P., & Rathi, Y. (2016). Joint multi-fiber NODDI parameter estimation and Tractography using the unscented information filter. *Frontiers in Neuroscience*, 10, 166.
- Roy, Maggie, Stephen Cunnane, Étienne Croteau, Alexandre Castellano, Mélanie Fortier, Félix C. Morency, Jean-Christophe Houde, and Maxime Descoteaux. 2017. "A Combined Dual-Tracer PET/diffusion Tractometry Analysis of the Posterior Cingulum in a Mild Cognitive Impairment Ketogenic Intervention." In *Proceedings of the International Society of Magnetic Resonance in Medicine (ISMRM)*. <http://scil.dinf.usherbrooke.ca/wp-content/papers/roy-et-al-ismrm18.pdf>.
- Schumacher, L. V., Reiser, M., Nitschke, K., Egger, K., Urbach, H., Hennig, J., ... Kaller, C. P. (2018). Probing the reproducibility of quantitative estimates of structural connectivity derived from global Tractography. *NeuroImage*, 175, 215–229.
- Shany, E., Inder, T. E., Goshen, S., Lee, I., Neil, J. J., Smyser, C. D., ... Shimony, J. S. (2017). Diffusion tensor Tractography of the cerebellar peduncles in prematurely born 7-year-old children. *Cerebellum*, 16(2), 314–325.
- Siless, V., Chang, K., Fischl, B., & Yendiki, A. (2018). Anatomical Cuts: Hierarchical clustering of Tractography streamlines based on anatomical similarity. *NeuroImage*, 166(February), 32–45.
- Sinke, M. R. T., Otte, W. M., Christiaens, D., Schmitt, O., Leemans, A., van der Toorn, A., ... Dijkhuizen, R. M. (2018). Diffusion MRI-based cortical connectome reconstruction: Dependency on Tractography procedures and neuroanatomical characteristics. *Brain Structure & Function*, 223(5), 2269–2285.
- Smith, R. E., Tournier, J.-D., Calamante, F., & Connelly, A. (2015). The effects of SIFT on the reproducibility and biological accuracy of the structural connectome. *NeuroImage*, 104, 253–265.
- Sporns, O., Tononi, G., & Kötter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Computational Biology*, 1(4), e42.
- Sydnor, V. J., Rivas-Grajales, A. M., Lyall, A. E., Zhang, F., Bouix, S., Karmacharya, S., et al. (2018). A comparison of three fiber tract delineation methods and their impact on White matter analysis. *NeuroImage*, 178(September), 318–331.
- Tensaouti, F., Lahlou, I., Clarisse, P., Lotterie, J. A., & Berry, I. (2011). Quantitative and reproducibility study of four Tractography algorithms used in clinical routine. *Journal of Magnetic Resonance Imaging: JMIRI*, 34(1), 165–172.
- Thomas, C., Ye, F. Q., Okan Irfanoglu, M., Modi, P., Saleem, K. S., Leopold, D. A., & Pierpaoli, C. (2014). Anatomical accuracy of brain connections derived from diffusion MRI Tractography is inherently limited. *Proceedings of the National Academy of Sciences of the United States of America*, 111(46), 16574–16579.
- Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., et al. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *NeuroImage*, 145(Pt B), 389–408.
- Vaessen, M. J., Hofman, P. A. M., Tijssen, H. N., Aldenkamp, A. P., Jansen, J. F. A., & Backes, W. H. (2010). The effect and reproducibility of different clinical DTI gradient sets on small world brain connectivity measures. *NeuroImage*, 51(3), 1106–1116.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, 80, 62–79.
- Visser, E., Nijhuis, E. H. J., Buitelaar, J. K., & Zwiers, M. P. (2011). Partition-based mass clustering of Tractography streamlines. *NeuroImage*, 54(1), 303–312.
- Vollmar, C., O'Muircheartaigh, J., Barker, G. J., Symms, M. R., Thompson, P., Kumari, V., ... Koepp, M. J. (2010). Identical, but not the same: Intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0T scanners. *NeuroImage*, 51(4), 1384–1394.
- Wakana, S., Caprihan, A., Panzenboeck, M. M., Fallon, J. H., Perry, M., Gollub, R. L., ... Mori, S. (2007). Reproducibility of quantitative Tractography methods applied to cerebral White matter. *NeuroImage*, 36(3), 630–644.
- Wang, J. Y., Abdi, H., Bakhadirov, K., Diaz-Arrastia, R., & Devous, M. D. (2012). A comprehensive reliability assessment of quantitative diffusion tensor Tractography. *NeuroImage*, 60(2), 1127–1138.
- Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., & Deriche, R. (2010). Unsupervised White matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for White matter fibers. *NeuroImage*, 51(1), 228–241.
- Wassermann, D., Makris, N., Rathi, Y., Shenton, M., Kikinis, R., Kubicki, M., & Westin, C.-F. (2016). The White matter query language: A novel approach for describing human White matter anatomy. *Brain Structure & Function*, 221(9), 4705–4721.
- Wu, C.-H., Hwang, T.-J., Chen, Y.-J., Hsu, Y.-C., Lo, Y.-C., Liu, C.-M., ... Isaac Tseng, W.-Y. (2015). Altered integrity of the right arcuate fasciculus as a trait marker of schizophrenia: A sibling study using Tractography-based analysis of the whole brain. *Human Brain Mapping*, 36(3), 1065–1076.
- Wu, Y., Zhang, F., Makris, N., Ning, Y., Norton, I., She, S., ... O'Donnell, L. J. (2018). Investigation into local White matter abnormality in emotional processing and sensorimotor areas using an automatically annotated fiber clustering in major depressive disorder. *NeuroImage*, 181(July), 16–29.
- Xia, Yan, U. Turken, Susan L. Whitfield-Gabrieli, and John D. Gabrieli. 2005. Knowledge-based classification of neuronal fibers in entire brain. *Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* (Vol. 8, pp. 205–212).
- Yamada, K., Sakai, K., Akazawa, K., Yuen, S., & Nishimura, T. (2009). MR Tractography: A review of its clinical applications. *Magnetic Resonance in Medical Sciences: MRMS: An Official Journal of Japan Society of Magnetic Resonance in Medicine*, 8(4), 165–174.
- Yeh, F.-C., Badre, D., & Verstynen, T. (2016). Connectometry: A statistical approach harnessing the analytical potential of the local connectome. *NeuroImage*, 125(January), 162–171.
- Yeh, F.-C., Panesar, S., Fernandes, D., Meola, A., Yoshino, M., Fernandez-Miranda, J. C., ... Verstynen, T. (2018). Population-averaged atlas of the macroscale human structural connectome and its network topology. *NeuroImage*, 178(September), 57–68.
- Yendiki, A., Reuter, M., Paul, W., Diana Rosas, H., & Fischl, B. (2016). Joint reconstruction of White-matter pathways from longitudinal diffusion MRI data with anatomical priors. *NeuroImage*, 127(February), 277–286.

- Yeo, S. S., Jang, S. H., & Son, S. M. (2014). The different maturation of the corticospinal tract and Corticoreticular pathway in Normal brain development: Diffusion tensor imaging study. *Frontiers in Human Neuroscience*, 8, 573.
- Zalesky, A., Cocchi, L., Fornito, A., Murray, M. M., & Bullmore, E. (2012). Connectivity differences in brain networks. *NeuroImage*, 60(2), 1055–1062.
- Zhang, F., Savadjiev, P., Cai, W., Y, S., Rathi, Y., Tunç, B., et al. (2018). Whole brain White matter connectivity analysis using machine learning: An application to autism. *NeuroImage*, 172(May), 826–837.
- Zhang, F., Wu, W., Ning, L., McAnulty, G., Waber, D., Gagoski, B., ... O'Donnell, L. J. (2018). Suprathreshold fiber cluster statistics: Leveraging White matter geometry to enhance Tractography statistical analysis. *NeuroImage*, 171(May), 341–354.
- Zhang, F., Wu, Y., Norton, I., Rigolo, L., Rathi, Y., Makris, N., & O'Donnell, L. J. (2018). An anatomically curated fiber clustering White matter atlas for consistent White matter tract Parcellation across the lifespan. *NeuroImage*, 179(October), 429–447.
- Zhang, F., P. Kahali, Y. Suter, I. Norton, L. Rigolo, P. Savadjiev, Y. Song, et al. 2017. Automated connectivity-based groupwise cortical atlas generation: Application to data of neurosurgical patients with brain tumors for cortical parcellation prediction. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 774–77.
- Zhang, F., I. Norton, W. Cai, Y. Song, W. M. Wells, and L. J. O'Donnell. 2017. Comparison between two white matter segmentation strategies: An investigation into white matter segmentation consistency. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 796–99.
- Zhang, Y., Zhang, J., Oishi, K., Faria, A. V., Jiang, H., Li, X., ... Mori, S. (2010). Atlas-guided tract reconstruction for automated and comprehensive examination of the White matter anatomy. *NeuroImage*, 52(4), 1289–1301.
- Zhang, Z., Descoteaux, M., Zhang, J., Girard, G., Chamberland, M., Dunson, D., ... Zhu, H. (2018). Mapping population-based structural connectomes. *NeuroImage*, 172, 130–145.
- Zhao, T., Duan, F., Liao, X., Dai, Z., Cao, M., He, Y., & Shu, N. (2015). Test-retest reliability of White matter structural brain networks: A multiband diffusion MRI study. *Frontiers in Human Neuroscience*, 9, 59.
- Ziyan, U., Sabuncu, M. R., Eric, W., Grimson, L., & Westin, C.-F. (2009). Consistency clustering: A robust algorithm for group-wise registration, segmentation and automatic atlas construction in diffusion MRI. *International Journal of Computer Vision*, 85(3), 279–290.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang F, Wu Y, Norton I, Rathi Y, Golby AJ, O'Donnell LJ. Test-retest reproducibility of white matter parcellation using diffusion MRI tractography fiber clustering. *Hum Brain Mapp.* 2019;40:3041–3057. <https://doi.org/10.1002/hbm.24579>

APPENDIX: DATA PREPROCESSING AND UKF TRACTOGRAPHY

For the HCP dataset, we used the already processed dMRI data (following the processing pipeline in (Glasser et al., 2013)). We extracted the $b = 3,000 \text{ s/mm}^2$ shell of 90 gradient directions and all b_0 scans (18) for

each subject, as applied in our previous studies (O'Donnell et al., 2017; Zhang, Kahali, et al., 2017; Zhang, Norton, et al., 2017; Zhang, Wu, Norton, et al., 2018). Angular resolution is better and more accurate at high b -values such as 3,000 (Descoteaux, Angelino, Fitzgibbons, & Deriche, 2007; Ning et al., 2015), and this single shell was chosen for reasonable computation time and memory use when performing tractography. The Freesurfer (version 5.2 was used in the HCP data processing pipeline) segmentation, which had been co-registered to the dMRI space, was directly used.

For the ABIDE-II and the PPMI datasets, we pre-processed the provided raw imaging data using the following steps. DWIConvert (<https://github.com/BRAINSia/BRAINStools>) was first applied to convert the original data format (DICOM or NIFTI) to NRRD. Eddy current-induced distortion correction and motion correction were conducted using the Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library tool (version 5.0.6) (Jenkinson et al., 2012). To further correct for distortions caused by magnetic field inhomogeneity (which leads to intensity loss and voxel shifts), an EPI distortion correction was performed with reference to the T2-weighted image using the ANTS (Avants et al., 2009). Because T2-weighted images were not available in all of these datasets (no T2 images were provided in the ABIDE-II, and in the PPMI dataset not all subjects had T2 images), we generated a synthetic T2-weighted image from a T1-weighted image for each subject (T1-weighted images were available for in all datasets) using the T1-weighted to T2-weighted conversion toolbox (<https://github.com/pnlbwh/T1toT2conversion>). For each subject, a nonlinear registration (registration was restricted to the phase encoding direction) was computed from the b_0 image to the synthetic T2-weighted image to make an EPI corrective warp. Then, the warp was applied to each diffusion image. A semi-automated quality control (using in-house developed Matlab scripts) was conducted on all diffusion images. Individuals that had diffusion images with any apparent signal drops were excluded from the analyses. For the remaining subjects, all gradient directions were retained for analysis. We also performed a Freesurfer (version 5.3) segmentation for each subject in these two datasets. Each individual's Freesurfer segmentation was transformed from T1-weighted space into diffusion corrected (b_0) space via nonlinear registration using ANTS.

After obtaining the pre-processed DWI data, we applied the same UKF parameters for all subjects under study, as follows. Tractography was seeded in all voxels within the brain mask where FA was greater than 0.1. Tracking stopped where the FA value fell below 0.08 or the normalized mean signal (the sum of the normalized signal across all gradient directions) fell below 0.06. The normalized average signal measure was employed to robustly distinguish between white/Gy matter and cerebrospinal fluid (CSF) regions. These seeding and stopping thresholds were set slightly below the default values (as used in our previous study (Zhang, Wu, Norton, et al., 2018)) to enable higher sensitivity for fiber tracking, in particular for the subjects (such as children) that might have low white matter anisotropy. Fibers that were longer than 40 mm were retained to avoid any bias toward implausible short fibers (Guevara et al., 2012; Jin et al., 2014; Lefranc et al., 2016).