



HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 October 01.

Published in final edited form as:

Neuroimage. 2021 October 01; 239: 118300. doi:10.1016/j.neuroimage.2021.118300.

Diffusion MRI and anatomic tracing in the same brain reveal common failure modes of tractography

Giorgia Grisot^a, Suzanne N. Haber^{b,c}, Anastasia Yendiki^{d,*}

^aDeepHealth, Inc., Belmont, MA, United States

^bDepartment of Pharmacology and Physiology, University of Rochester, Rochester, NY, United States

^cMcLean Hospital, Belmont, MA, United States

^dAthinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA, United States

Abstract

Anatomic tracing is recognized as a critical source of knowledge on brain circuitry that can be used to assess the accuracy of diffusion MRI (dMRI) tractography. However, most prior studies that have performed such assessments have used dMRI and tracer data from different brains and/or have been limited in the scope of dMRI analysis methods allowed by the data. In this work, we perform a quantitative, voxel-wise comparison of dMRI tractography and anatomic tracing data in the same macaque brain. An ex vivo dMRI acquisition with high angular resolution and high maximum b-value allows us to compare a range of q-space sampling, orientation reconstruction, and tractography strategies. The availability of tracing in the same brain allows us to localize the sources of tractography errors and to identify axonal configurations that lead to such errors consistently, across dMRI acquisition and analysis strategies. We find that these common failure modes involve geometries such as branching or turning, which cannot be modeled well by crossing fibers. We also find that the default thresholds that are commonly used in tractography correspond to rather conservative, low-sensitivity operating points. While deterministic tractography tends to have higher sensitivity than probabilistic tractography in that very conservative threshold regime, the latter outperforms the former as the threshold is relaxed to avoid missing true anatomical connections. On the other hand, the q-space sampling scheme and maximum b-value have less of an impact on accuracy. Finally, using scans from a set of additional macaque brains, we show that there is enough inter-individual variability to warrant caution when dMRI and tracer data come from different animals, as is often the case in the tractography validation literature. Taken together, our results provide insights on the limitations of

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

*Corresponding author. ayendiki@mgh.harvard.edu (A. Yendiki).

Data available Statements

Data used in this paper are available through the IronTract challenge. The reconstruction and tractography methods are publicly available.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118300.

current tractography methods and on the critical role that anatomic tracing can play in identifying potential avenues for improvement.

Keywords

Diffusion MRI; Validation; Tractography; Tracing

1. Introduction

Diffusion MRI (dMRI) is the only available tool for studying the structural connections of the brain non-invasively and in vivo. It allows us to measure the probability distribution of water molecule displacements in tissue (Le Bihan et al. 1986), and hence to infer the orientation of the underlying microstructure (Moseley et al. 1990; Chenevert et al., 1990; Basser and Pierpaoli 1996). In white matter, this provides estimates of the local orientations of axon bundles. Tractography algorithms follow these orientation vectors from voxel to voxel and attempt to reconstruct the trajectories of white-matter pathways through the brain (Parker et al. 2002; Koch et al., 2002; Fillard et al., 2009; Kreher et al., 2008; Mori et al. 1999). Tractography is essential for characterizing brain networks (Hagmann et al., 2008; Bullmore and Sporns 2009) and, in combination with other microstructural properties derived from dMRI, has the potential to further our understanding of a plethora of neurological and psychiatric conditions.

The potential of dMRI to provide unique information about the connective anatomy of the brain is hampered by its limited resolution. In the human brain, the diameter of axons is on the order of 1–3 μm , whereas a voxel in an in vivo dMRI scan is, at best, 1–2 mm wide. As a result, hundreds of thousands of axons may contribute to the diffusion signal of a single voxel. These axons can be organized in many possible configurations, and there are different configurations that can give rise to a very similar diffusion profile. As a result, the inference of anatomical connections from the orientation distributions that are typically reconstructed from dMRI is an ill-posed problem.

The reconstruction of white-matter pathways from dMRI data involves several methodological choices. These include the data acquisition scheme (e.g., single-shell vs. multi-shell vs. full sampling of q-space on a Cartesian grid), the orientation reconstruction method (e.g., tensor vs. crossing-fiber models or model-based vs. model-free), and the tractography approach (e.g., deterministic vs. probabilistic). Given that both the acquisition time and the computational costs of these strategies vary widely, it is important to know whether the more demanding approaches lead to substantial gains in terms of the anatomical accuracy of the reconstructed pathways.

Answering this question requires the quantitative assessment of the accuracy of different data acquisition and analysis strategies. This is not straightforward because it is difficult to establish the true connective anatomy of an individual brain. One may assess relatively easily whether certain connections, which are known to exist from the anatomical literature, are present or absent in the output of tractography, but not whether any single voxel that tractography has labeled as part of a given connection is accurate or not. As a result, early

validation studies of tractography in the human brain were mostly qualitative, involving visual comparison to gross anatomical dissections (Lawes et al. 2008; Martino et al. 2011).

Quantitative assessment of tractography accuracy is possible in phantoms, which are objects with known fiber configurations (Fieremans et al. 2008; Fillard et al., 2011; Perrin et al. 2005; Poupon et al. 2008; Daducci et al., 2014; Leemans et al. 2005; Neher et al. 2014). Phantoms may be defined digitally or they may be physical objects built with artificial fibers. In the former case, dMRI data are simulated by applying a forward model to the digital phantom; in the latter, data are acquired by imaging the physical phantom in an MRI scanner. In either case, phantoms involve rather simple configurations of a small set of fiber bundles, which do not capture the full complexity of white-matter architecture. Although they are useful for benchmarking tractography methods, the performance of a specific method in this setting cannot predict its performance in reconstructing any given fiber pathway in the brain.

Validation of tractography can be performed in animal brains using MR-visible tracers (Knösche et al. 2015; Dyrby et al. 2007; Leergaard et al. 2003; Yamada et al. 2008; Lin et al. 2003; Lin et al. 2001). After such a tracer is injected into the brain and allowed to propagate along fiber bundles, its concentration can be imaged with MRI. However, these tracers are not guaranteed to follow the entire trajectory of all axon bundles from the injection site all the way to their terminals, missing smaller or more diffuse projections. Furthermore, the quality of the acquired tracer images is limited by the resolution and signal-to-noise ratio (SNR) of MRI.

Anatomic tracing is the only technique that can provide full reconstructions of the axons projecting to or from the point of injection (Jbabdi et al. 2015; Lehman et al. 2011). Several studies have compared tractography to anatomic tracing in monkeys, either qualitatively (Schmahmann et al. 2007; Jbabdi et al. 2013; Calabrese et al. 2014; Safadi et al. 2018) or quantitatively (Dauguet et al. 2007; Hagmann et al., 2008; Gao et al. 2013; Thomas et al. 2014; van den Heuvel et al. 2015; Azadbakht et al. 2015; Donahue et al. 2016; Schilling, Gao, et al. 2019; Tang et al. 2019; Ambrosen et al. 2020; Girard et al., 2020). The quantitative studies had at least one of the following limitations: *(i)* They did not have anatomic tracing and dMRI tractography available in the same brains (Hagmann et al., 2008; Thomas et al. 2014; van den Heuvel et al. 2015; Azadbakht et al. 2015; Donahue et al. 2016; Ambrosen et al. 2020; Girard et al., 2020), or *(ii)* Their quantitative comparisons involved only the terminals and not the full trajectory of the axons (Hagmann et al., 2008; Gao et al. 2013; van den Heuvel et al. 2015; Azadbakht et al. 2015; Donahue et al. 2016; Tang et al. 2019; Ambrosen et al. 2020; Girard et al., 2020), or *(iii)* Their dMRI acquisitions were limited either by a low b-value or by low spatial or angular resolution (Dauguet et al. 2007; Gao et al. 2013; van den Heuvel et al. 2015; Schilling, Gao, et al. 2019). Finally, most prior quantitative studies used dMRI data with a single b-value, thus precluding the use of analysis methods that require multi-shell or full sampling of q-space.

The majority of these prior studies compared tracer experiments to dMRI scans acquired from different animals. However, the impact of this on the resulting estimates of tractography accuracy has not been quantified. The inter-individual variability of geometric

features, such as the shape of cortical foldings or subcortical structures, is lower in the macaque than the human brain. Although accurate alignment of these features across animals is an easier problem to solve, it does not guarantee alignment of axon bundles. Indeed, tractography can be sensitive to even small shifts in the seeding area. Thus, it is important to investigate the impact of using tracer and dMRI data from different animals.

We have recently acquired a set of high-resolution, high-SNR, ex vivo dMRI scans from macaque brains that have also received tracer injections (Safadi et al. 2018; Tang et al. 2019). In previous work, we used these data to investigate how projections of the prefrontal cortex are organized within the anterior limb of the internal capsule (Safadi et al. 2018). Specifically, we demonstrated that the relative positions of these axon bundles with respect to each other, as seen in the tracer data, can be replicated with dMRI tractography in both macaques and humans. We also used this data set to show that tractography can replicate, in both macaques and humans, a connectional hub in the rostral anterior cingulate cortex that is identified by tracer injections (Tang et al. 2019). These studies focused on specific connectivity questions and hence used a single tractography method. Here we use this data set to address methodological questions regarding the accuracy of different diffusion sampling, orientation reconstruction, and tractography strategies.

Importantly, our tracer experiments and dMRI scans were carried out in the same animal. The tracer experiments allow us to follow the axons as they leave the injection site, branch into different fiber bundles, and travel to their terminal fields. Thus, we can trace the entire trajectory of different groups of axons from a single point. The advantage of having the dMRI in the same animal is that it allows us to validate this entire trajectory of axon bundles, rather than just the terminations of the axons. As a result, we can find not only *if* tractography goes wrong but exactly *where* it goes wrong. Furthermore, the dMRI scan protocol allows us to compare a range of acquisition and analysis strategies.

Among prior validation studies, those that compared the accuracy of multiple tractography methods used either numerically simulated data (Côté et al. 2013; Maier-Hein et al. 2017) or anatomic tracing (Gao et al. 2013; Thomas et al. 2014; Schilling, Gao, et al. 2019, et al. 2019). The present work differs from those studies in several ways. First, those studies used dMRI data with a single, low to moderate b-value. At present, however, the state of the art has shifted away from such acquisitions, with several high-profile, large-scale human studies now using multi-shell (Miller et al. 2016; Casey et al. 2018; Somerville et al. 2018) or undersampled Cartesian grid (Tobisch et al. 2018) acquisition schemes for dMRI. Here we compare the accuracy of single-shell, multi-shell, and Cartesian schemes. Second, previous studies compared different tractography methods (e.g., deterministic vs. probabilistic) using default thresholds, which result in different sensitivity *and* different specificity for each type of method. Here we probe the sensitivity-specificity trade-off by varying the tractography thresholds. This allows us to evaluate the sensitivity of different methods at the same level of specificity, making side-by-side comparisons intuitive.

We focus on the following questions: 1) How does the choice of q-space sampling scheme, orientation reconstruction method, and tractography method contribute to the accuracy of the reconstructed pathways? 2) What types of axonal configurations cause tractography errors

consistently, across acquisition and analysis strategies? 3) Do our conclusions change when we use dMRI and tracer data from the same or different brains? The results herein expand on the analyses presented in (Grisot et al., 2018).

2. Methods

2.1. Overview of experimental design

The reference data that we use to assess the accuracy of dMRI tractography in this work are manually labeled axon bundles from a tracer injection in the frontopolar cortex of a single macaque brain, referred to as MAC1. First, we apply various orientation reconstruction and tractography methods to a dMRI scan of MAC1, and we assess the accuracy of each method by comparing its output to anatomic tracing in the same brain. We then apply the same analysis methods to dMRI scans of 13 different macaque brains (MAC2–14) and compare them to the tracer data from MAC1. This allows us to investigate whether we would reach the same conclusions regarding the performance of different dMRI analysis methods if we compared anatomic tracing to dMRI tractography from the same or different brains.

2.2. Anatomical tracer injection

Surgery and tissue preparation were performed at the University of Rochester Medical Center. Details of these procedures were described previously (Lehman et al. 2011; Safadi et al. 2018). Briefly, an adult male monkey (*Macaca Mulatta*) received an injection of the bidirectional tracer Lucifer Yellow (LY) conjugated with dextran amine (40–50 nl, 10% in 0.1 M phosphate buffer, pH 7.4; Invitrogen) in the frontopolar cortex (Brodmann area 10). Twelve days after the injection, the brain was removed, postfixed overnight and cryoprotected in increasing gradients of sucrose (10, 20, and 30%). All experiments were performed in accordance with the Institute of Laboratory Animal Resources Guide for the Care and Use of Laboratory Animals and approved by the University of Rochester Committee on Animal Resources. The animal came with full health records and underwent physical and behavioral screening to ensure the absence of neuropathy. In addition, histological sections were examined to ascertain that there was no gliosis or other damage.

This injection in the frontopolar cortex was part of a larger collection aimed at mapping the circuitry of the frontal, prefrontal, and cingulate cortex (Safadi et al., 2018; Tang et al., 2019). The injection was selected for this validation study as it combined high-quality tracer data and a dMRI scan in the same brain. Furthermore, as described in the following, the tracing revealed a complex system of projections, with multiple junction areas that proved challenging for tractography.

2.3. MRI acquisition

After fixation, each brain was shipped to the Athinoula A. Martinos Center for Biomedical Imaging for MRI scanning. The scan was performed in a small-bore 4.7 T Bruker BioSpin MRI system, with an internal gradient diameter of 116 mm, maximum gradient strength 480 mT/m, and birdcage volume RF coil internal diameter of 72 mm. During scanning, the brain was submerged in liquid Fomblin (Solvay Solexis Inc.) to eliminate susceptibility artifacts at air-tissue interfaces. We used a two-shot, 3D echo-planar imaging (EPI) dMRI

sequence with TR=750 ms, TE=43 ms, matrix size $96 \times 96 \times 112$, and 0.7 mm isotropic resolution. We collected 1 non-diffusion weighted ($b = 0$) and 514 diffusion-weighted volumes, corresponding to a cubic lattice in q-space contained within the interior of a ball of maximum radius $b_{\max}=40,000$ s/mm², with $\delta=15$ ms and $\Delta=19$ ms. The total acquisition time was 48 h. Given that diffusivity in fixed tissue is approximately a fourth of that observed in vivo (Dyrby et al. 2011), our b_{\max} value was roughly equivalent to $b_{\max}=10,000$ s/mm² in vivo. While the diffusion times used here are somewhat different than those used for high-b dMRI in vivo elsewhere (Fan et al. 2016), they still yield a diffusion length scale that is both greater than axon diameters ($\sim 1 \mu\text{m}$, see Aboitiz et al. 1992) and smaller than the axon undulation wavelength ($\sim 20 \mu\text{m}$, see Fontana 1781, Lee et al. 2020), ensuring statistical independence of diffusion transverse and parallel to axons. Furthermore, we have previously used similar diffusion times as in the present study to reconstruct diffusion orientations in *post mortem* brain and validated them against optical imaging (Jones et al. 2020).

The dMRI scan described above was performed on MAC1 (the brain that had received the frontopolar tracer injection described above), as well as the brains of 13 additional adult male macaques (which had received other injections that are not used in the present work). We used the *eddy* tool in FSL to compensate for eddy-current distortions (Andersson and Sotiropoulos, 2016). The segmented acquisition mitigated EPI distortions, hence no further correction for such distortions was performed in post-processing. The SNR, defined as the mean over the standard deviation of the $b = 0$ signal over the brain area traversed by the tracer, was 5.25 for MAC1 and 5.24 ± 1.15 on average for the full set of macaques. Thus, MAC1 was representative of the average quality of the set.

2.4. Histological processing

After MRI scanning, the brain was shipped back to the University of Rochester for histological processing. Serial coronal sections of $50 \mu\text{m}$ were cut on a freezing microtome. Before each section was cut, the undistorted blockface was photographed, for use during registration to the ex vivo dMRI data. Immunocytochemistry was performed on every 8th slice to visualize the transported tracer, resulting in an inter-slice resolution of $400 \mu\text{m}$. Before incubation in primary antisera, tissue was treated with 10% methanol and 3% hydrogen peroxide in 0.1 M PB to inhibit endogenous peroxidase activity, rinsed in PB with 0.3% Triton X-100 (TX; Sigma), and preincubated in 10% normal goat serum (NGS) and 0.3% TX in PB for 30 min. Tissue was placed in the primary anti-LY (1:3000 dilution; Invitrogen) in 10% NGS and 0.3% TX in PB for 4 nights at 4°C . After extensive rinsing, the tissue was incubated in biotinylated secondary antibody, followed by incubation with the avidin–biotin complex solution (Vectastain ABC kit; Vector Laboratories). Immunoreactivity was visualized using standard 3,3-diaminobenzidine tetra-hydrochloride (DAB) procedures. Staining was intensified by incubating the tissue for 5–15 min in a solution of 0.05% DAB, 0.025% cobalt chloride, 0.02% nickel ammonium sulfate, and 0.01% H_2O_2 to yield a black reaction product. Sections were mounted onto gelcoated slides, dehydrated, defatted in xylenes, and cover-slipped with Permount (Haber et al. 2006; Lehman et al. 2011; Haynes and Haber 2013).

Labeled fiber bundles were outlined under dark-field illumination with a 4.0 or 6.4x objective, using Neurolucida software (MBF Bioscience). Fibers traveling together were outlined as a group or bundle. Axons were charted as they left the tracer injection site and followed through the right hemisphere, until the anterior commissure. Fiber orientation within a bundle was indicated by charting individual fibers within the bundle outline (Fig. 1A). These fiber orientations were used as a visual aid in the process of outlining the bundles manually across sections (Fig. 1A–1B). Key features within each section, including gray and white matter boundaries, as well as major subcortical structures, were outlined to evaluate the subsequent registration across sections. The 2D outlines were combined across slices using IMOD software (Boulder Laboratory; Kremer et al., 1996) to create 3D renderings of the structures and pathways as they traveled through them (Fig. 1B). These were used to further refine bundle contours and ensure spatial consistency across sections.

2.5. MRI-to-histology registration

We aligned the histological sections and dMRI data of MAC1 as follows (Fig. 1C). **1) Histology to blockface.** Each histology slide was registered to its corresponding blockface in two steps. The first step used a robust affine registration that detects outlier areas, i.e., areas that do not have direct correspondence between the two images, and gives these areas less weight when computing the registration (Reuter et al., 2010). The second step used a symmetric diffeomorphic optimizer that maximizes cross-correlation of images within the space of diffeomorphic maps (B. B. Avants et al. 2008). Histology-to-blockface registration is a very common practice intended to compensate for distortions due to histological processing (Yushkevich et al. 2006; Majka and Wójcik, 2016). It ensures one-to-one correspondence between each distorted histological slice and an undistorted photograph of the same slice taken before cutting it. Thus, 2D registration can be performed between the two, reducing the degrees of freedom. **2) Blockface to dMRI.** All blockface images were stacked to create a 3D volume and registered to the $b = 0$ dMRI volume using a 3D affine registration followed by a 3D diffeomorphic registration, using the same registration tools as the previous step.

The transformations obtained from the above steps were then applied to the tracing outlines, to transfer them from the space of the distorted histology slides to the space of the dMRI scan. The same transformations were applied to map the location of the injection site from histology to dMRI space, to use it as the seed region for tractography.

We also aligned the dMRI data from MAC1, the brain that had received the tracer injection, to the dMRI data from each of the other 13 macaque brains with the combination of affine and non-linear registration tools described above. We used the resulting warp field to map tracings and injection masks from MAC1 to every other macaque's native dMRI space.

2.6. Q-space resampling

Our goal was to validate dMRI tractography across several q-space sampling schemes. To this end, we used the dMRI data of each monkey, which had been acquired on a Cartesian grid in q-space, to generate three additional datasets with different q-space sampling schemes: (i) a Cartesian grid with a reduced q-space field of view, (ii) a single q-shell, and

(iii) 3 q-shells. We obtained the first one by extracting the 257 lower-q diffusion-weighted volumes (Grid-25.6 K) from the original dataset (Grid-40 K), resulting in a Cartesian grid with $b_{\max}=25,600$ s/mm². The generation of the q-shell datasets required approximating data points in q-space distributed on a sphere from data points distributed on a Cartesian grid. We did this via the non-uniform fast Fourier transform (Fessler and Sutton 2003), an approach to q-space resampling that we presented and validated previously (Jones et al. 2020). In the present work, we used this method to generate a multi-shell dataset that comprised 64, 64, and 128 directions, respectively, with $b = 4000, 8000, \text{ and } 12000$ s/mm². These b-values are approximately equivalent to in vivo b-values of 1000, 2000, and 3000 s/mm² (Dyrby et al. 2011). We also generated a single-shell data set that comprised the 128-direction, $b = 12000$ s/mm² data only. For each shell, we selected gradient vectors that were uniformly distributed on the sphere (Caruyer et al. 2013).

2.7. Diffusion orientation reconstruction

We applied six methods for estimating the orientations of axonal bundles from the dMRI data: diffusion tensor imaging (DTI), ball-and-stick (BS), generalized q-space imaging (GQI), diffusion spectrum imaging (DSI), q-ball imaging (QBI), and QBI with constant solid angle (QBI-CSA). A brief description of each method that we used in this work, as well as the software that it is implemented in, is included below. Unless otherwise noted, all reconstructions were performed using the default parameters provided by the respective software package.

DTI: The tensor model assumes that diffusion follows a Gaussian distribution (Basser et al., 1994). Given its inability to resolve crossing fibers, we included DTI here to establish a lower bound for the performance of the remaining methods. We used the DSI Studio toolbox (<http://dsi-studio.labsolver.org>) to perform least-squares tensor fitting.

BS: This model decomposes the dMRI signal of a voxel into an isotropic compartment (“ball”) and multiple anisotropic compartments (“sticks”) that represent the principal fiber populations in that voxel (Behrens et al. 2007). Model fitting was performed with BedpostX (Behrens et al. 2003; Behrens et al. 2007) (part of FSL software), which uses Markov chain Monte Carlo sampling to generate probability distributions of the BS model parameters. An extension of the BS model that uses a Gamma distribution of diffusivities to better fit the non-monoexponential decay of the signal and reduce fiber orientation over-fitting (Jbabdi et al. 2012) was applied to the datasets with more than one b-value. We fit a maximum of three anisotropic compartments per voxel.

DSI: DSI is a model-free technique that takes advantage of the Fourier relationship between the MR signal in q-space and the diffusion propagator (Wedeen et al. 2005). The orientation distribution function (ODF) in each voxel is then computed by radial integration of the diffusion propagator. To be able to retrieve the ODF by means of a Fourier transform, the diffusion data needs to be acquired with a Cartesian grid sampling scheme. DSI reconstruction was performed with DSI Studio.

GQI: GQI is another model-free method that quantifies the density of diffusing water at different orientations (Yeh et al., 2010). It can be applied to dMRI data acquired with arbitrary sampling schemes, including grid, single-shell, and multi-shell. Reconstruction was performed with DSI Studio, using a diffusion length ratio of 0.57 and a maximum of three fiber orientations per voxel.

QBI: This approach uses the Funk-Radon transform of the diffusion measurements on a single q-shell to approximate the ODF in each voxel (Tuch 2004). Fiber orientations are estimated as the local maxima of the ODFs. Qboot (Sotiropoulos et al. 2011) (part of the FSL software) was used for reconstruction, with a maximum of three fiber orientations per voxel. The ODF shape and the probability distributions of the fiber orientations were obtained using residual bootstrapping (Whitcher et al. 2008).

QBI-CSA: This adaptation of the original QBI approach provides a mathematically correct formulation of the ODF by taking into account the quadratic growth of the volume element in the radial dimension, leading to sharper ODFs (Aganj et al. 2010). Moreover, ODF estimation is performed using the spherical harmonics basis. QBI-CSA reconstruction was also performed using Qboot.

Table 1 shows all the feasible combinations of sampling scheme and orientation reconstruction approach. We analyzed the Cartesian-grid datasets (Grid-40 K and Grid-25.6 K) with DTI, BS, GQI and DSI; the multi-shell datasets with DTI, BS, GQI, and QBI-CSA; and the single-shell datasets with DTI, BS, GQI, QBI-CSA and QBI.

2.8. Diffusion tractography

We used the injection site in the frontal pole, extended into the superficial white matter, as the seed region for tractography. We performed deterministic tractography with a generalized version of the Fiber Assignment by Continuous Tracking (FACT) algorithm that uses quantitative anisotropy as the termination index (Yeh et al. 2013), as implemented in DSI Studio. We performed probabilistic tractography using FSL's probtrackx algorithm, which models the orientation of a fiber population as a distribution and, at each step, draws a sample from it and progresses along the sampled orientation vector (Behrens et al. 2007). For BS, QBI, and QBI-CSA, sample directions from these distributions were automatically generated during the modeling stage by FSL. To perform probabilistic tractography with GQI and DSI, we drew samples from the ODFs reconstructed with GQI and DSI. We followed an approach similar to (Tournier et al., 2012), where a distribution of fiber orientations around the ODF peaks is generated by performing a random selection of the ODF vertices weighted by their amplitude. Samples are, therefore, more likely to be drawn from orientations where the ODF amplitude is large.

2.9. Comparison of tractography and tracing

We compared the output of each tractography method to the tracings that we had manually charted in MAC1 and mapped to the dMRI space of each brain (Fig. 2). A voxel reached by dMRI tractography was deemed a true positive (TP) if it was also reached by the tracer, and a false positive (FP) otherwise. For each combination of diffusion sampling

scheme, orientation estimation method, and tractography algorithm, we produced a receiver-operating characteristic (ROC) curve, by plotting the true positive rate (TRP) vs. the false positive rate (FPR). The TPR (sensitivity) is defined as the fraction of voxels reached by the tracer that were also reached by tractography streamlines, and the FPR (1-specificity) is defined as the fraction of voxels without tracing that were reached by tractography streamlines. Such plots are commonly used to evaluate the performance of a binary classifier. We obtained different points on the ROC curve by varying common user-defined parameters: the threshold of the voxel visitation map for probabilistic tractography, and the bending angle threshold for deterministic tractography. We also considered varying the fractional anisotropy (FA) threshold for deterministic tractography, but this did not yield as wide a range of TP and FP values as the bending angle threshold (results not shown).

We compared the performance of each combination of diffusion sampling scheme, orientation estimation method, and tractography algorithm at three typical operating points along the ROC curve. Each of these three points was defined by selecting a threshold for one method, computing the specificity of that method at the selected threshold, and then setting the threshold of each of the other methods to achieve the same specificity. In more detail, the selected thresholds were: 1) Default deterministic threshold (T_{det}): We set the angle threshold of deterministic tensor tractography to the commonly used value of 40° , and found its specificity at that threshold. 2) Default probabilistic threshold (T_{prob}): We set the probability threshold of probabilistic BS tractography to the commonly used value of 0.01, and found its specificity at that threshold. 3) Anatomically defined threshold (T_{anat}): We set the probability threshold of probabilistic BS tractography to ensure that the tractography reached the corpus callosum (CC), internal capsule (IC), and external capsule (EC), and found its specificity at that threshold. (The CC, IC, and EC were among the main white-matter bundles that axons from the injection site traveled through, as revealed by the tracing data.) After finding the operating points where each dMRI analysis approach achieved the three levels of specificity described above, we compared their sensitivity at those three levels of specificity.

We carried out two types of comparisons. First, we compared the tracing to dMRI tractography in MAC1, the brain that had received the tracer injection. Second, we compared the tracing from MAC1 to dMRI tractography from the other 13 brains. This allowed us to assess whether we would reach the same conclusions if the tracing data came from a different or the same brain as the dMRI data. We wanted to gauge whether differences observed between the results that we obtained from the dMRI data of MAC1 and those that we obtained from the other brains could simply be explained by imaging noise, in which case they would have also been observed between two dMRI data sets of the same brain. To this end, we compared the inter- and intra-individual variability of our single-shell data sets. For each macaque brain, we split the 128 directions of the single-shell dataset into two subsets of 64 evenly distributed gradient directions, which we refer to as subsets 1 and 2. We then processed each subset of directions in the same fashion as the original single-shell dataset and computed their sensitivity and the same levels of specificity (T_{det} , T_{prob} , and T_{anat}). We quantified intra-individual variability by comparing the sensitivity of tractography between the subset 1 and subset 2 data sets of the same brain. We quantified inter-individual variability by comparing the sensitivity of tractography between the subset 1 dataset of

MAC1 and the subset 2 data set of each of the remaining 13 brains. We computed averages of the intra- and inter-individual sensitivity differences across all brains.

3. Results

3.1. Comparison of anatomic tracing and dMRI tractography from the same animal

Fig. 3 summarizes the results of the ROC analysis for the animal with the tracer injection. The green and red shaded areas contain, respectively, the ROC curves for probabilistic and deterministic tractography, with all combinations of reconstruction method and q-space sampling scheme. Supplemental Figure S1 is a more detailed version of this figure, showing the individual ROC curves of all methods.

As seen in Fig. 3, the operating points corresponding to the default deterministic and probabilistic thresholds (T_{det} , T_{prob}) were rather conservative, with low FPR and low TPR. At those tractography thresholds, several of the main projections of the injection site, as revealed by the tracer data, would be missed by dMRI tractography. The operating point where these projections would be detected by tractography (T_{anat}) would require us to tolerate an FPR about 3 times as high as those of the default thresholds. At this higher FPR level of the anatomically defined threshold, there were greater differences in TPR between probabilistic and deterministic tractography methods. At the two default thresholds (T_{det} , T_{prob}), the different tractography methods performed more similarly to each other.

The area demarcated by the gray box in Fig. 3 is magnified in the plots of Fig. 4, where ROC curves are shown separately for each q-space sampling scheme. Each color represents a different orientation reconstruction method. Curves with and without square markers show results, respectively, from probabilistic and deterministic tractography. Supplemental Figure S2 shows the same curves grouped by orientation reconstruction method. Table 2 lists the sensitivity of each method at the three operating points: T_{det} , T_{prob} , and T_{anat} .

3.1.1. Deterministic vs. probabilistic tractography—The differences between deterministic and probabilistic methods became more accentuated as sensitivity increased. Across reconstruction methods and q-space sampling schemes, probabilistic tractography achieved a greater increase in sensitivity than deterministic tractography, for the same loss of specificity. At the higher-sensitivity, anatomically defined threshold (T_{anat}), all reconstruction methods performed better with probabilistic than deterministic tractography (see Table 2). At the more conservative, low-sensitivity operating points that correspond to the default thresholds (T_{det} , T_{prob}), deterministic tractography exhibited, for the most part, higher sensitivity than probabilistic tractography. This was the case even for the BS model, which is typically used with probabilistic tractography. The exceptions were multi- and single-shell GQI, as well as full-grid DSI, which performed better with probabilistic than deterministic tractography, even at the conservative, default thresholds.

3.1.2. Comparison of orientation reconstruction methods—As seen in Table 2 and Supplemental Figure S1, probabilistic GQI, particularly when applied to multi- or single-shell data, performed better than all other reconstruction methods, across all thresholds. This is an interesting finding, as GQI has heretofore been used mostly with

deterministic tractography. Other methods that performed quite well were deterministic BS or QBI-CSA at the lower-sensitivity operating points, and probabilistic QBI-CSA at the higher-sensitivity operating points.

3.1.3. Comparison of q-space sampling schemes—As seen in Supplemental Figure S2, when probabilistic tractography was used, there was little difference in the performance of multi- vs. single-shell data, or in the performance of full- vs. reduced-b grid data. Differences between sampling schemes were more pronounced when deterministic tractography was used, with multi-shell data outperforming single-shell data. This was also the case for probabilistic BS tractography, the only type of probabilistic tractography that showed such an effect.

3.1.4. Model fitting error analysis—In the absence of ground truth, a common figure-of-merit for evaluating model-based diffusion orientation reconstruction methods is the residual error of the model fit. For example, we expect the tensor model to fit single-shell data better than data collected with multiple b-values, and the BS model to fit single- and multi-shell data better than Cartesian grid data. However, it is not clear whether lower residual error translates to higher anatomical accuracy in tractography. We investigated this by computing the residual error of these models for each q-space sampling scheme. We found that the BS model fit the resampled shell data better than it fit the original, Cartesian-grid dataset, with root-mean-squared errors (RMSE) below 8% (Supplemental Figure S3–a). However, a lower RMSE did not necessarily imply better anatomical accuracy. For example, the BS model exhibited similar accuracy on data sampled on a single shell or a grid in q-space (see Supplemental Figure S2), despite the fact that it fit the former better than the latter.

Similarly, we compared the ODF-based reconstruction methods by assessing how well they can reproduce the ODF obtained from a DSI reconstruction, which we used as a reference. For all ODFs, we found high correlation coefficients with the DSI ODF (Supplemental Figure S3–b). Again, however, higher correlation with the reference ODF did not necessarily translate to better tractography accuracy. For example, QBI and QBI-CSA exhibited similar accuracy on single-shell data (Fig. 4), despite the fact that ODFs reconstructed with QBI were much more highly correlated to the reference ODFs than those reconstructed with QBI-CSA.

3.1.5. Where does tractography go wrong?—An important benefit of collecting tracer and dMRI data on the same brain is that it allows us to compare the two on a voxel-by-voxel basis, and thus to identify the exact locations where tractography errors occur. This can provide valuable information on the types of axonal configurations that pose a challenge for tractography and thus point towards possible directions for future methodological improvements.

Fig. 5 shows photomicrographs of representative histological sections, where the tracer is visualized under darkfield microscopy. Fibers from an injection site exit the gray matter in a single compact bundle, referred to as a “stalk” (Krieg 1973). Here, the stalk contains fibers that travel tightly bundled and then split into CC, IC, EC, and striatal bundles. Other

axons exit the injection site, travel at the edges of the stalk for a short distance, and then fan out, either to project to nearby areas of cortex or to merge into white-matter bundles that reach distant cortical areas. For this particular injection site, groups of axons in this category either fanned out to terminate in the dorsomedial and dorsolateral prefrontal cortex (PFC) or entered the superior longitudinal fasciculus (SLF) and uncinate fasciculus (UF).

We identified locations where errors occurred consistently across tractography methods by obtaining a histogram of TPs. For each combination of q-space sampling, orientation reconstruction and tractography strategy, we extracted a map of TP voxels at a specified FPR level. We summed these binary maps across all methods in the MAC1 brain. Figs. 6a and 6b show this histogram of TPs at two FPR levels, corresponding to the default deterministic threshold T_{det} and anatomically defined threshold T_{anat} , respectively. Each histogram is displayed as a maximum intensity projection (maximum value across axial slices), which allows us to summarize the contents of multiple slices in a single 2D view. Fig. 6c shows a 3D isosurface of the tracer injection, where we mark four examples of junction areas that proved challenging for tractography. These were areas where fibers: 1. Fan out towards the lateral PFC; 2. Branch off the stalk and enter the IC/EC; 3. Turn into the UF; and 4. Follow the CC and turn towards the contralateral hemisphere. These four areas are examined in more detail in Figs. 7–10.

In Figs. 6a and 6b, areas where tractography errors are common can be identified as areas with abrupt changes in the number of tractography methods that achieve TPs. When tractography operates at low FP levels (Fig. 6a), it often fails at following fibers through the aforementioned junction areas where axons fan, branch into multiple bundles, or take a sharp turn. This is unsurprising, as, in the absence of any information on which ODF peak to follow, tractography algorithms are configured to minimize their bending angle and travel as straight as possible. This captures crossing fiber configurations better than it does fanning, branching, and turning. When tractography thresholds are relaxed (Fig. 6b), more methods are able to find the previously missing bundles. In particular, the CC, IC, and EC branches are identified by the majority of methods at the T_{anat} threshold. However, this happens at the expense of increased FPs (which are not shown in this figure). Furthermore, other branches continue to be missed, even at this threshold.

The four trouble areas from Fig. 6c are examined further in Figs. 7–10. We illustrate common errors by showing single slices from the histograms of TPs (Figs. 7–9) or FPs (Fig. 10) across all combinations of q-space sampling, orientation reconstruction, and tractography strategies. We juxtapose these maps to photomicrographs of the same slices that show the fibers from the tracer injection under darkfield microscopy.

1. **Fanning area: lateral PFC** (Fig. 7). These fibers are among the first ones to separate from the dense stalk projecting from the injection site. As they fan out, they keep traveling caudally almost parallel to the EC. As seen in Fig. 7, all tractography methods missed a substantial portion of these lateral PFC fibers (green), even at the more relaxed T_{anat} threshold. In most cases, tractography streamlines followed the denser group of fibers in the main stalk (inset i) and disregarded the sparser group of fibers fanning out to the lateral PFC (inset ii).

2. **Branching area: IC/EC** (Fig. 8). The IC and EC were, along with the CC, the branches that most tractography methods were able to reconstruct correctly. This, however, required a tractography threshold more liberal than the default (compare T_{anat} vs. T_{det}).
3. **Turning area: UF** (Fig. 9). The tracer injection identified a group of fibers from the frontal pole that traveled toward the temporal cortex via the UF. These fibers traveled caudally in the EC until the anterior commissure, where they turned laterally, leaving the EC and joining the UF. Furthermore, as the tracing in Fig. 9 shows, after turning into the UF, fibers turned sharply in the dorso-rostral direction and into the middle longitudinal fasciculus. This double turn was particularly challenging for tractography. Most streamlines continued to follow the EC caudally, instead of turning towards the UF. Of the streamlines that did turn into the UF correctly, none completed the double turn at the thresholds shown here. In some cases, it may be possible to reconstruct this turn correctly by relaxing the threshold even more (e.g., for deterministic methods, by using a bending angle threshold above 85°), but this would come with an excessive increase in FPs. Note that all crossing-fiber reconstruction methods considered here were able to recover a peak pointing in the direction of the UF in that location. Hence any problems with capturing the turn into the UF correctly would not be solved by making the ODF peaks more accurate or sharper; they can only be solved by being able to differentiate between a crossing and a turn, which conventional crossing-fiber methods cannot do.
4. **Turning area: CC** (Fig. 10). Here we show an example of an FP that was observed consistently across tractography methods. This error is related to a sharp turn of CC fibers. The tracer data show fibers traveling in the ventral-most region of the genu of the CC, right at the interface between callosal white matter and the septum pellucidum (see location 4 in Fig. 6c). Instead of remaining in the CC, which has a high curvature at the location where it turns towards the contralateral hemisphere, tractography streamlines continue erroneously in a rostro-caudal direction, through the septum pellucidum and the fornix. As seen in Fig. 10, this FP occurs with some methods even at a very stringent threshold (T_{det}) and it occurs with even more methods at the threshold that is necessary to capture the true projections through the EC, IC, and CC (T_{anat}). Thus, removing the FP by thresholding would also eliminate some of these true projections. Furthermore, because this FP occurs consistently across methods, majority voting among multiple tractography algorithms would not eliminate it, either.

3.2. Comparison of anatomic tracing and dMRI tractography from different animals

Figs. 11 and 12 show how the results of this validation study would change if we used tracer data from a different brain than the one that received the dMRI scan, as is often done in the literature. We show results at T_{det} and T_{anat} , which are representative, respectively, of lower- and higher-sensitivity operating points. Results at T_{prob} were very similar to those at T_{det} , hence they are omitted in this section.

Fig. 11 compares the intra- vs. inter-individual variability of sensitivity at the same level of specificity. The plots show sensitivity differences between single-shell dMRI datasets extracted from a scan of the same brain (averaged over all 14 brains), and between a single-shell dataset from MAC1 and from each of the other 13 brains. As seen in the figure, the differences observed when the tracer injection was compared to dMRI tractography in the same vs. different animals (red and orange) were much greater than the differences between two dMRI data sets from the same brain (blue and green). This suggests that any inconsistencies in the findings from the brain with the tracer injection and the other brains are unlikely to be simply due to imaging noise.

Note that the methods with the lowest TPR (like DTI) have the most reliable TPR and methods with the highest TPR (like GQI) have the least reliable TPR. As seen in Fig. 11, this difference in test-retest reliability is more pronounced at a very stringent threshold (T_{det}) and becomes less pronounced at a more relaxed threshold (T_{anat}). The TPR of the lowest-performing methods plateaus at a relatively stringent threshold and does not increase substantially as the threshold is relaxed (see Fig. 4). The TPR of the highest-performing methods, however, increases sharply at first and only plateaus at a more relaxed threshold. This sharper increase in TPR, at the low-TPR operating range, makes these methods more sensitive to the threshold in that range, which likely contributes to their lower test-retest reliability in that threshold range. Once the performance of a method plateaus, its TPR becomes more reliable.

Fig. 12 compares, for all the dMRI sampling schemes and reconstruction methods that we considered, the sensitivity of dMRI tractography as measured in the macaque with the tracer injection (stars) and the average sensitivity measured in the other 13 animals (circles with standard error bars). In some cases, the sensitivity obtained when the tracer injection and dMRI data come from the same brain overlaps with the range of sensitivities obtained by comparing the tracer data to dMRI tractography in different brains. In many cases, however, using dMRI and tracer data from different brains leads to an over- or under-estimation of sensitivity. Some of the patterns that we observed by comparing dMRI and anatomic tracing in MAC1, such as the greater performance differences between tractography methods at the anatomically defined threshold than the default threshold, or the higher sensitivity of probabilistic tractography at the former threshold, could also be observed when comparing the tracer injection from MAC1 to dMRI tractography in the other 13 macaques. However, the variability introduced by using dMRI and tracer data from different brains was substantial enough to potentially confound the comparison between methods.

Supplemental Figure S4 shows how the outcome of a comparison between probabilistic and deterministic tractography would vary, depending on the dMRI dataset used to assess their accuracy. Each subplot in Supplemental Figure S4 compares the TPR of probabilistic and deterministic tractography for a different orientation reconstruction method and q-space sampling scheme. Each open circle in these plots is the outcome of comparing the tracer injection from MAC1 to dMRI tractography in one of the other 13 macaques. The filled circle shows the average sensitivity from these 13 cases. The star is the outcome of comparing the tracer injection and dMRI tractography from MAC1.

Supplemental Figure S5 shows a similar comparison between q-space sampling schemes. The first and second columns of plots compare the grid sampling scheme with $b_{\max}=40$ K vs. $b_{\max}=25.6$ K. The third and fourth columns compare the multi-shell vs. single-shell sampling scheme. Each color represents a different dMRI orientation reconstruction method.

The results in Supplemental Figures S4–S5 suggest that, *on average*, the outcome of comparing tracer and dMRI data from different brains often agrees with the outcome of comparing tracer and dMRI data from the same brain. That is, the group-averaged sensitivities replicated some of the findings of the comparison of tracing and tractography in MAC1: deterministic tractography performs better in the conservative, low-sensitivity operating point (T_{det}); probabilistic tractography performs better at the anatomically defined operating point (T_{anat}), which was required to capture the main bundles that the injection site projects to; probabilistic GQI was the top-performing method; the $b_{\max}=40$ K and 25.6 K grid sampling schemes performed very similarly; multi-shell and single-shell also performed similarly, with the former having an advantage over the latter mostly when deterministic tractography was used.

Importantly, however, although on average over the 13 brains we were able to replicate the main findings from the comparison in the brain with the injection, there was substantial variability among individual cases. Therefore, when the comparison was performed on the basis of dMRI data from a single brain that differed from the brain with the injection, the outcome often disagreed with the comparison of dMRI and tracing in the same brain. These discrepancies often exceeded the ~5% intra-individual differences seen in Fig. 11, thus they are unlikely to be explained by imaging noise alone. Comparing tractography methods on the basis of individual tracing and dMRI datasets from different brains is common in the literature. Our results indicate that such an approach may introduce confounds due to inter-individual variability.

4. Discussion

This study used ex vivo dMRI scans of macaque brains acquired with high angular and spatial resolution, as well as a dense sampling of q-space. These data allowed us to evaluate a wider range of q-space sampling schemes, beyond the single-shell data used by other validation studies that performed quantitative comparisons of multiple tractography methods (Côté et al. 2013; Gao et al. 2013; Thomas et al. 2014; Maier-Hein et al. 2017; Schilling, Gao, et al. 2019, et al. 2019). We performed a systematic evaluation of different q-space sampling, orientation reconstruction, and tractography strategies, by comparing their output to a tracer injection in the frontopolar cortex of the same brain. In addition to investigating the differences between methods, the availability of tracer data in the same brain allowed us to localize the sources of errors that occurred consistently across different acquisition and analysis strategies, and thus to identify common failures modes of dMRI tractography. Data from the present study are also used for the IronTract challenge (<https://irontract.mgh.harvard.edu>), which will be an opportunity to test an even wider range of analytical tools.

4.1. Probabilistic vs. deterministic tractography

As seen in the ROC curves of Fig. 3, the choice between probabilistic and deterministic streamline propagation emerged as a factor with a major impact on the accuracy of tractography. Previous validation studies that compared deterministic and probabilistic methods used either numerically simulated data (Côté et al. 2013; Maier-Hein et al. 2017) or anatomic tracing (Gao et al. 2013; Thomas et al. 2014; Schilling, Gao, et al. 2019, et al. 2019). A finding often reported in this literature is that deterministic tractography tends to produce fewer incorrect connections, but that probabilistic tractography tends to produce more complete reconstructions of the correct connections. This result, however, reflects the fact that deterministic and probabilistic methods were compared at their respective default thresholds, which are characterized by *both different specificity and different sensitivity*. Therefore, it is more reflective of the default thresholds of these methods, rather than the methods themselves.

Given that, in principle, one may always increase the specificity of a classifier by decreasing its sensitivity and vice versa, the classical way of determining which of two classifiers has superior performance is by comparing their *sensitivity at the same level of specificity*. This can be accomplished by varying the threshold of each classifier and tracing its ROC curve, as we have done here. Our results suggest that the default thresholds for both deterministic and probabilistic tractography (T_{det} , T_{prob}) are quite conservative, operating at low-sensitivity, high-specificity points on the ROC curve. This may make sense in the general use case of tractography, where we lack prior anatomical information on what brain regions each seed region is connected to. With these thresholds, however, there is a high rate of false negatives, i.e., tractography misses several true anatomical connections. Recovering the main projections of the injection site through the CC, IC, and EC (Fig. 8) required relaxing the threshold to a less conservative level (T_{anat}), with roughly 3 times the FPR of the default thresholds. Even at this less conservative threshold, tractography missed some of the more challenging bundles, such as the projections to the lateral PFC (Fig. 7) and UF (Fig. 9).

At very conservative thresholds (very low-sensitivity, high-specificity points on the ROC curve), deterministic tractography appears to have an advantage. At the more relaxed thresholds that are required to capture the main projections of the injection site (T_{anat} and beyond), probabilistic tractography has a definitive advantage, achieving much higher sensitivity than deterministic tractography at the same specificity level. Two recent studies that performed full ROC analyses also found probabilistic tractography to have higher sensitivity than deterministic tractography at the same level of specificity (Delettre et al., 2019; Girard et al., 2020). Those studies, however, compared area-to-area connectivity matrices between tractography and databases of tracer data from a different set of animals. That is, they evaluated the accuracy of the cortical terminations and not the full trajectory of tractography.

Ultimately, a way to get around the sensitivity/specificity trade-off is to use prior anatomical information: choose a very liberal threshold, operating at very high sensitivity and low specificity and thus ensuring that all true connections are included, and then use anatomical regions of interest to remove the false connections. While this is a valid (and in fact very

common) way of deploying tractography, it amounts to using it as a tool for segmenting known brain connections rather than a tool for discovery.

4.2. Comparison of q-space sampling schemes

As mentioned above, previous validation studies that compared multiple orientation reconstruction and tractography methods used mainly single-shell data. A prior comparison of tractography with single-shell vs. full-grid sampling in the macaque brain (Calabrese et al. 2014) kept acquisition time constant across sampling schemes by using a higher spatial resolution for the former than the latter. This makes it difficult to disentangle the effects of the q-space sampling scheme on tractography accuracy from those of the spatial resolution. A recent validation study that used BS probabilistic tractography on single-shell data with different b-values, and compared tractography terminals to an existing database of macaque tracer injections, found that the effect of the b-value was not statistically significant (Ambrosen et al. 2020). A study that did not validate tractography but diffusion orientations, by comparing them to fiber orientations from histological sections, also found that increasing the b-value or number of directions of single-shell data had diminishing returns (Schilling et al., 2018).

In some cases, the increase in diffusion resolution afforded by sampling high-q data points may be offset by the higher noise levels in those data points. For example, we sometimes found the 257-volume DSI scheme to outperform the 515-volume DSI scheme. A study that evaluated the angular error of 203- and 515-volume DSI with respect to an even more densely sampled reference scan, found that the 203-volume scheme performed slightly better on in vivo data (Kuo et al., 2008). Under-sampled DSI reconstructed with compressed sensing has also been shown to give lower errors than its fully sampled counterpart, when compared to a high-SNR, fully sampled reference scan (Bilgic et al., 2013). This suggests that collecting fewer data points and interpolating the rest may have denoising effects. Similarly, extrapolating the tails of the diffusion propagator from lower-q, higher-SNR data points may have some benefits over measuring the high-q, low-SNR points directly.

Overall, the four q-space sampling schemes that we evaluated here achieved a similar range of performance, as seen in Fig. 4. This is despite the fact that the reduced Cartesian grid and shell sampling schemes had, respectively, 0.67 and 0.3 times the maximum b-value of the full Cartesian grid. This finding is consistent with a recent study where we compared ODF peaks from same sampling schemes to axonal orientations measured with optical imaging in post mortem human brain samples (Jones et al., 2020). In that work we found that, by choosing the reconstruction method appropriately, it was possible to achieve similar accuracy with single-shell as with multi-shell data, and to come within 10% of the accuracy of the full Cartesian-grid acquisition.

These results may seem disappointing and perhaps counterintuitive, given the current trends in hardware and sequence design towards collecting more and higher b-values in vivo. Indeed, more gradient directions are expected improve the angular resolution, reducing the minimum angle between crossing fiber bundles that can be resolved. Higher b-values are expected to sharpen ODF peaks, again allowing better modeling of crossing bundles. However, it is important to recall that the common failure modes that we identified are not

due to crossings but due to other fiber configurations, like branching and turning. These errors would not be resolved by improved modeling of crossings. Taking better advantage of the information contained in data with high b-values and high angular resolution to resolve such ambiguities is an open problem, and it may require reconstruction methods that go beyond the crossing-fiber paradigm (Yendiki et al., 2020). Furthermore, it may be worth revisiting methods that attempt to resolve certain types of fanning configurations, either by modeling (Savadjiev et al., 2008) or by using information from neighboring voxels (Reisert et al., 2012; Bastiani et al., 2017).

4.3. Comparison of orientation reconstruction methods

Although we found performance differences between reconstruction methods, these differences were not as pronounced as the ones between deterministic and probabilistic tractography. The top-performing reconstruction method was GQI combined with probabilistic tractography. This is contrary to the conventional use of GQI, which has heretofore been used with deterministic tractography (Yeh et al., 2010). The high performance of GQI, which yields less sharp ODFs than QBI or other related methods, confirms that there is no simple relationship between sharper ODFs and more accurate tractography. While a sharper ODF would be better at modeling crossing fibers, a less sharp ODF may be better at capturing other fiber geometries, such as fanning or branching, particularly if it is combined with a probabilistic technique that samples a spectrum of orientations around each ODF peak.

Another figure-of-merit that is often used to assess the quality of model-based dMRI reconstruction methods, when ground truth is not available, is the residual error that expresses how well the model fits the dMRI signal (Jbabdi et al. 2012). However, we found that lower residual error does not necessarily translate to greater anatomical accuracy (Supplemental Figure S3).

4.4. Localization of tractography errors

In the majority of studies that validate dMRI tractography against tracer injection data, the latter come from existing databases that only include information on the end points of each connection (Hagmann et al., 2008; Gao et al. 2013; van den Heuvel et al. 2015; Azadbakht et al. 2015; Donahue et al. 2016; Ambrosen et al. 2020). These studies can only assess the accuracy of tractography based on the end points of the tracts, i.e., based on a connectivity matrix. This approach can tell us if a tractography algorithm missed a true connection between two brain regions or if it detected a false connection between regions, but it cannot tell us where along the path between the two regions the error(s) occurred.

The tracer data used in this work included the trajectories of axon bundles through the white matter, rather than termination points only. The availability of such data, in the same brain as the dMRI scan, allowed us to assess the accuracy of tractography on a voxel-by-voxel basis and to identify locations where errors occurred consistently, across dMRI acquisition, reconstruction, and tractography strategies. We found that these were often areas that contained fiber configurations like branching or turning. This illustrates the importance of tracer data for identifying realistic failure modes of tractography that

go beyond the idealized “crossing vs. kissing” configurations previously used in digital or physical phantoms.

Branching, turning, or fanning fibers are not handled well by conventional, ODF-based methods, which have been designed to model crossing fibers. Even when the ODFs have multiple peaks that can capture the different branches or the different sections of a turn, tractography algorithms are designed to follow a trajectory that minimizes curvature, as would be consistent with a crossing. In other work, where we used optical imaging to validate diffusion orientations in human brain samples, we showed that increasing the spatial resolution of dMRI may help reduce errors due to branching fibers (Jones et al. 2020). Another way of overcoming this ambiguity of crossing-fiber models is to include prior anatomical information in the tractography algorithm. While tracer injection data cannot be mapped directly from the macaque to the human brain, we have shown that organizational rules on the relative positions of axon bundles with respect to each other generalize across species (Safadi et al. 2018; Jbabdi et al. 2013). Such information on organizational rules could be incorporated, for example, in a tractography method that uses prior information on the relative positions of white-matter bundles with respect to the surrounding anatomy (Yendiki et al. 2011).

4.5. Comparison of anatomic tracing and dMRI tractography from different animals

Anatomic studies show that general organizational principles of connective anatomy are reproducible across individuals and across species. Thus, high-level information on the terminations or the route followed by brain connections (e.g., “the frontal pole projects to the lateral PFC” or “the frontal pole projects through the IC”) can be transferred across studies. Caution is needed, however, when performing a voxel-wise assessment of the accuracy of tractography. Previously we have shown that, although the relative positions of small axon bundles with respect to each other are consistent across individual human subjects, their absolute positions do not necessarily agree (Safadi et al. 2018).

Here we investigated inter- vs. intra-individual variability in the context of using tracer data from the same macaque brain as the dMRI data or tracer data transferred from a different macaque brain. Note that the results presented here are likely to represent a best-case scenario in terms of the alignment across animals. That is because we registered the tracer data of MAC1 to the dMRI data of MAC1, and then used the latter to register to the dMRI data of other animals. In a typical situation where dMRI data from MAC1 would not be available, one would have to register the tracer data of one animal to the dMRI data of another. This across-modality *and* across-individual registration would be even more challenging.

Our results show that certain broad findings, such as the relative merits of probabilistic and deterministic tractography at different thresholds, or the similar performance of different q-space sampling schemes, could be replicated on average when using dMRI data from a set of brains other than the one with the tracer injection (Supplemental Figures S4–S5). However, there was enough inter-individual variability to caution against drawing conclusions from single cases where the tracer data have been transferred from one brain to another. This variability was greater than the intra-individual variability observed when using different

dMRI data sets from the same brain (Figs. 11 and 12), hence it is unlikely to be explained solely by imaging noise in the dMRI data.

Using dMRI and tracer data from different animals may be more appropriate for some types of comparisons than others. Specifically, there are three levels of granularity at which dMRI tractography and tracer data can be compared: *(i)* Comparison of connected areas (a.k.a., the “connectivity matrix” approach). At this coarse level, tractography is deemed to be correct when it finds a connection between a pair of brain regions that are known to be connected based on tracer experiments, and incorrect otherwise. Given that these general connection patterns are highly reproducible between individual brains, it is reasonable to perform such a comparison using dMRI data and tracer experiments from different brains. *(ii)* Comparison of topographies. At this intermediate level of granularity, tractography is deemed to be correct when it reproduces the topographic organization of axon bundles within the large white-matter pathways, as revealed by tracing. We have previously shown that these organizational rules are highly similar across individual brains and even across the macaque and human brain (Safadi et al. 2018). Thus, such a comparison can also be performed between dMRI and tracer experiments carried out in different brains. *(iii)* Voxel-by-voxel comparison. At this much finer level, tractography is compared to tracing in terms of the precise route of axon bundles through the brain. While comparisons of dMRI and tracing at the coarser levels (*i - ii*) are valuable for assessing *how often* tractography errors occur, voxel-wise comparisons (*iii*) are the only way to determine exactly *where* the errors occur. This is the type of comparison that we have performed here. Even if the general patterns of connective anatomy are similar across brains, it is unlikely that image registration will lead to perfect voxel-wise alignment of all bundles, and especially of the small groups of axons that are labeled by a tracer injection. Our results demonstrate that a voxel-wise comparison of dMRI tractography and tracer data from the same brain is necessary to accurately identify and explore how errors occur.

4.6. Limitations

A limitation of this study is that we evaluated the accuracy of tractography for a single cortical injection site. Preliminary results from an open tractography challenge using these data suggest that optimizing the parameters of tractography methods for one injection site does not always yield optimal settings for a different injection site (Maffei et al., 2020). It would not be surprising if, for example, the relative ranking of different dMRI reconstruction methods or the optimal parameters of each method varied across seed areas. However, we expect that general principles, such as the types of fiber geometries that cause systematic errors, the effects of using probabilistic vs. deterministic tractography, *etc.*, will be applicable broadly. This is something to be investigated further in future work.

Tracer data may suffer from various limitations, and experience with tracer studies is key to ensure data quality. The tracer may be taken up by fibers of passage and the exact area of axonal uptake at the injection site can be difficult to determine. Inconsistency in uptake and transport may result in variable quality between injections. The injection used in this study passed rigorous quality assurance checks at Dr. Haber’s laboratory and had high-quality transport. The manual annotation of the axon bundles was also checked by Dr. Haber and

refined at multiple stages. Two other experiments in which injections had been placed at this site, but in different animals (not included here), served as controls.

As is the case for all methods that rely on histological processing, distortions due to sectioning and staining can interfere with the alignment of histological sections. With a technique like a myelin stain, which would label all myelinated axons within a section, it would be very difficult to ensure that axon bundles are aligned across sections. However, tracer studies have the advantage that only axons that are connected to the injection site are labeled by the tracer. This makes it is easier to check the alignment of the labeled axon bundles between consecutive sections.

Both the tracing maps and the tractography voxel visitation maps that we used were binarized, i.e., we only took into account whether tracing or tractography went through a certain voxel or not, and did not use any information on the relative density of axon or streamline bundles, respectively. Obtaining this information from the tracer data would be valuable, as there is substantial variability in the density of different axon bundles projecting from the same injection site (see Figs. 7–9). This would require labeling individual axons in the tracer data in the future.

Finally, each of the q-space points in our dMRI data was sampled only once. This precludes the use of cross-validation to evaluate the variability of our results for all the sampling schemes that we have studied here. The results of Section 3.2, where we subdivided the single-shell data into two subsets, provide some measure of this variability. However, this point should be investigated more extensively by acquiring multiple repetitions of each scan in future studies.

4.7. Alternative approaches

Although most of our knowledge on brain circuitry comes from anatomic tracer studies, there are several other techniques for visualizing axons in post mortem brains at microscopic resolution. Techniques that have been used for post mortem validation of fiber orientations obtained from dMRI include traditional histological methods such as myelin stains (Leergaard et al., 2010; Choe et al., 2012; Seehaus et al., 2015) and DiI (Schilling et al., 2016, 2018), label-free optical imaging such polarization microscopy (Mollink, 2017) and optical coherence tomography (Jones et al., 2020), or tissue clearing followed by staining with fluorescent dyes (Leuze et al., 2021). For a review of these techniques, their merits, and their limitations, we refer the reader to Yendiki et al. (2021).

5. Conclusions

We have performed a systematic evaluation of different q-space sampling, orientation reconstruction, and tractography strategies by quantifying their sensitivity and specificity in reconstructing connections identified with anatomic tracing. The availability of dMRI and tracer data from the same macaque brains enable a quantitative, voxel-wise assessment of tractography accuracy. This has allowed us to identify common failure modes of tractography across dMRI acquisition and analysis strategies, involving axonal configurations that cannot be captured by crossing-fiber models. Our findings illustrate the

importance of datasets with dMRI and tracing in the same brains for gaining insights into the fundamental limitations of tractography.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institute of Mental Health (R01-MH045573, P50-MH106435). Additional research support was provided by the National Institute of Biomedical Imaging and Bioengineering (R01-EB021265) and the National Institute of Neurological Disorders and Stroke (R01-NS119911). Imaging was carried out at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41-EB015896, a P41 Biotechnology Resource Grant, and instrumentation supported by the NIH Shared Instrumentation Grant Program (S10RR016811, S10RR023401, S10RR019307, and S10RR023043).

References

- Aboitiz F, et al., 1992. Fiber composition of the human corpus callosum. *Brain Res*598 (1–2), 143–153. [PubMed: 1486477]
- Aganj I, et al., 2010. Reconstruction of the orientation distribution function in single- and multiple-shell q-ball imaging within constant solid angle. *Magn. Reson. Med.* 64 (2), 554–566. Available at <http://www.ncbi.nlm.nih.gov/pubmed/20535807>. [Accessed December 11, 2017]. [PubMed: 20535807]
- Ambrosen KS, et al., 2020. Validation of structural brain connectivity networks: the impact of scanning parameters. *Neuroimage*204, 116207. [PubMed: 31539592]
- Andersson JL, Sotiropoulos SN, 2016. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage*125, 1063–1078. [PubMed: 26481672]
- Avants BB, et al., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12 (1), 26–41. Available at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2276735&tool=pmcentrez&rendertype=abstract>. [Accessed July 16, 2014]. [PubMed: 17659998]
- Azadbakht H, et al., 2015. Validation of high-resolution tractography against in vivo tracing in the macaque visual cortex. *Cerebral Cortex (New York, N.Y. : 1991)*25 (11), 4299–4309. Available at <http://www.ncbi.nlm.nih.gov/pubmed/25787833>. [Accessed March 30, 2017].
- Basser PJ, Mattiello J, LeBihan D, 1994. MR diffusion tensor spectroscopy and imaging. *Biophys. J.* 66 (1), 259–267. [PubMed: 8130344]
- Basser PJ, Pierpaoli C, 1996. Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magnetic Resonance. Series B*111 (3), 209–219.
- Bastiani M, Cottaar M, Dikranian K, et al., 2017. Improved tractography using asymmetric fibre orientation distributions. *Neuroimage*158, 205–218. [PubMed: 28669902]
- Behrens TEJ, et al., 2003. Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magn. Reson. Med.* 50 (5), 1077–1088. Available at <http://www.ncbi.nlm.nih.gov/pubmed/14587019>. [Accessed April 4, 2017]. [PubMed: 14587019]
- Behrens TEJ, et al., 2007. Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage*34 (1), 144–155. [PubMed: 17070705]
- Le Bihan D, et al., 1986. MR imaging of intravoxel incoherent motions: application to diffusion and perfusion in neurologic disorders. *Radiology*161 (2), 401–407. Available at 10.1148/radiology.161.2.3763909. [Accessed February 18, 2015]. [PubMed: 3763909]
- Bilgic B, Chatnuntawech I, Setsompop K, Cauley SF, Yendiki A, Wald LL, Adalsteinsson E, 2013. Fast dictionary-based reconstruction for diffusion spectrum imaging. *IEEE Trans. Med. Imaging*32 (11), 2022–2033. [PubMed: 23846466]

- Bullmore E, Sporns O, 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10 (3), 186–198. Available at <http://www.nature.com/articles/nrn2575>. [Accessed April 1, 2018]. [PubMed: 19190637]
- Calabrese E, et al., 2014. Investigating the tradeoffs between spatial resolution and diffusion sampling for brain mapping with diffusion tractography: time well spent? *Hum Brain Mapp* 35 (11), 5667–5685. [PubMed: 25044786]
- Caruyer E, et al., 2013. Design of multishell sampling schemes with uniform coverage in diffusion MRI. *Magn. Reson. Med.* 69 (6), 1534–1540. Available at 10.1002/mrm.24736. [PubMed: 23625329]
- Casey BJ, et al., 2018. The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54. [PubMed: 29567376]
- Chenevert TL, Brunberg JA, Pipe JG, 1990. Anisotropic diffusion in human white matter: demonstration with MR techniques in vivo. *Radiology* 177 (2), 401–405. [PubMed: 2217776]
- Choe A, Stepniewska I, Colvin D, Ding Z, Anderson A, 2012. Validation of diffusion tensor MRI in the central nervous system using light microscopy: quantitative comparison of fiber properties. *NMR Biomed.* 25 (7), 900–908. [PubMed: 22246940]
- Côté MA, et al., 2013. Tractometer: towards validation of tractography pipelines. *Med. Image Anal.*
- Daducci A, Canales-Rodriguez EJ, Descoteaux M, Garyfallidis E, Gur Y, Lin Y–C, et al., 2014. Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI. *IEEE Trans. Med. Imaging* 33 (2), 384–399. Available at <http://www.ncbi.nlm.nih.gov/pubmed/24132007>. [Accessed February 28, 2018]. [PubMed: 24132007]
- Dauguet J, et al., 2007. Comparison of fiber tracts derived from in-vivo DTI tractography with 3D histological neural tract tracer reconstruction on a macaque brain. *Neuroimage* 37 (2), 530–538. Available at <http://www.sciencedirect.com/science/article/pii/S105381190700328X>. [Accessed April 13, 2015]. [PubMed: 17604650]
- Delettre C, Messé A, Dell LA, Foubet O, Heuer K, Larrat B, Meriaux S, Mangin JF, Reillo I, de Juan Romero C, Borrell V, Toro R, Hilgetag CC, 2019. Comparison between diffusion MRI tractography and histological tract-tracing of cortico-cortical structural connectivity in the ferret brain. *Netw Neurosci* 3 (4), 1038–1050. [PubMed: 31637337]
- Donahue CJ, et al., 2016. Using diffusion tractography to predict cortical connection strength and distance: a quantitative comparison with tracers in the monkey. *The J. Neurosci.* 36 (25), 6758–6770. Available at <http://www.ncbi.nlm.nih.gov/pubmed/27335406>. [Accessed March 30, 2017]. [PubMed: 27335406]
- Dyrby TB, et al., 2011. An ex vivo imaging pipeline for producing high-quality and high-resolution diffusion-weighted imaging datasets. *Hum. Brain Mapp.* 32 (4), 544–563. [PubMed: 20945352]
- Dyrby TB, et al., 2007. Validation of in vitro probabilistic tractography. *Neuroimage* 37 (4), 1267–1277. [PubMed: 17706434]
- Fan Q, et al., 2016. MGH-USC Human Connectome Project datasets with ultra-high b–value diffusion MRI. *Neuroimage* 124 (Pt B), 1108–1114. [PubMed: 26364861]
- Fessler JA, Sutton BP, 2003. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Trans. Signal Process.* 51 (2), 560–574.
- Fieremans E, et al., 2008. The design of anisotropic diffusion phantoms for the validation of diffusion weighted magnetic resonance imaging. *Phys. Med. Biol.* 53 (19), 5405–5419. Available at 10.1088/0031-9155/53/19/009. [PubMed: 18765890]
- Fillard P, et al., 2011. Quantitative evaluation of 10 tractography algorithms on a realistic diffusion MR phantom. *Neuroimage*.
- Fillard P, Poupon C, Mangin JF, 2009. A novel global tractography algorithm based on an adaptive spin glass model. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 927–934.
- Fontana F, 1781. *Traité sur le venin de la vipère sur les poisons américaines sur le laurier-cerise et sur quelques autres poisons végétaux*, vol. 2, chez Nyon l’Ainé.
- Gao Y, et al., 2013. Validation of DTI tractography-based measures of primary motor area connectivity in the squirrel monkey brain. *PLoS*

One8 (10), e75065. Available at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3788067&tool=pmcentrez&rendertype=abstract>. [Accessed October 20, 2014].

- Girard G, Caminiti R, Battaglia-Mayer A, St-Onge E, Ambrosen KS, Eskildsen SF, Krug K, Dyrby TB, Descoteaux M, Thiran JP, Innocenti GM, 2020. On the cortical connectivity in the macaque brain: a comparison of diffusion tractography and histological tracing data. *Neuroimage*221, 117201. [PubMed: 32739552]
- Grisot G, Haber SN, Yendiki A, 2018. Validation of diffusion MRI models and tractography algorithms using chemical tracing. *Proc. Intl. Soc. Mag. Res. Med.*
- Haber SN, et al., 2006. Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J. Neurosci.* 26, 8368–8376. [PubMed: 16899732]
- Hagmann P, Gigandet X, Meuli R, 2008. Quantitative validation of MR tractography using the CoCoMac database. *Proc. Intl. Soc. Mag. Reson. Med.* 16. Available at <http://infoscience.epfl.ch/record/135048/files/00427.pdf?version=1%5Cnpapers2://publication/uuid/20D8B169-24F9-44DA-891A-34E0B7A22548>.
- Haynes WIA, Haber SN, 2013. The organization of prefrontal-subthalamic inputs in primates provides an anatomical substrate for both functional specificity and integration: implications for basal ganglia models and deep brain stimulation. *J. Neurosci.* 33 (11), 4804–4814. Available at 10.1523/JNEUROSCI.4674-12.2013. [PubMed: 23486951]
- van den Heuvel MP, et al., 2015. Comparison of diffusion tractography and tract-tracing measures of connectivity strength in rhesus macaque connectome. *Hum. Brain Mapp.* 36 (8), 3064–3075. Available at 10.1002/hbm.22828. [Accessed March 30, 2017]. [PubMed: 26058702]
- Jbabdi S, et al., 2013. Human and monkey ventral prefrontal fibers use the same organizational principles to reach their targets: tracing versus tractography. *J. Neurosci.* 33 (7), 3190–3201. Available at <http://www.jneurosci.org/content/33/7/3190.long#ref-33>. [PubMed: 23407972]
- Jbabdi S, et al., 2015. Measuring macroscopic brain connections in vivo. *Nat. Neurosci.* 18 (11), 1546–1555. Available at <http://www.ncbi.nlm.nih.gov/pubmed/26505566>. [Accessed March 30, 2017]. [PubMed: 26505566]
- Jbabdi S, et al., 2012. Model-based analysis of multishell diffusion MR data for tractography: how to get over fitting problems. *Magn. Reson. Med.* 68 (6), 1846–1855. [PubMed: 22334356]
- Jones R, et al., 2020. Insight into the fundamental trade-offs of diffusion MRI from polarization-sensitive optical coherence tomography in ex vivo human brain. *Neuroimage*214, 116704. [PubMed: 32151760]
- Knösche TR, et al., 2015. Validation of tractography: comparison with manganese tracing. *Hum. Brain Mapp.* 36 (10), 4116–4134. Available at <http://www.ncbi.nlm.nih.gov/pubmed/26178765>. [Accessed March 30, 2017]. [PubMed: 26178765]
- Koch MA, Norris DG & Hund-Georgiadis M, 2002. An investigation of functional and anatomical connectivity using magnetic resonance imaging
- Kreher BW, Mader I, Kiselev VG, 2008. Gibbs tracking: a novel approach for the reconstruction of neuronal pathways. *Magn. Reson. Med.* 60 (4), 953–963. [PubMed: 18816816]
- Kremer JR, Mastrorade DN, McIntosh JR, 1996. Computer Visualization of Three-Dimensional Image Data Using IMOD. *J. Struct. Biol.* 116 (1), 71–76. Available at <http://www.ncbi.nlm.nih.gov/pubmed/8742726>. [Accessed April 7, 2017]. [PubMed: 8742726]
- Krieg W, 1973. *Architectonics of the Human Cerebral Fiber Systems*. Brain Books, Evanston, IL.
- Kuo LW, et al., 2008. Optimization of diffusion spectrum imaging and q-ball imaging on clinical MRI system. *Neuroimage*41 (1), 7–18. Available at <http://www.ncbi.nlm.nih.gov/pubmed/18387822>. [Accessed February 26, 2018]. [PubMed: 18387822]
- Lawes INC, et al., 2008. Atlas-based segmentation of white matter tracts of the human brain using diffusion tensor tractography and comparison with classical dissection. *Neuroimage*39 (1), 62–79. [PubMed: 17919935]
- Lee HH, et al., 2020. The impact of realistic axonal shape on axon diameter estimation using diffusion MRI. *Neuroimage*223, 117228.

- Leemans A, et al., 2005. Mathematical framework for simulating diffusion tensor MR neural fiber bundles. *Magn. Reson. Med.* 53 (4), 944–953. Available at 10.1002/mrm.20418. [Accessed April 1, 2018]. [PubMed: 15799061]
- Leergaard TB, et al., 2003. In vivo tracing of major rat brain pathways using manganese-enhanced magnetic resonance imaging and three-dimensional digital atlas. *Neuroimage* 20 (3), 1591–1600. [PubMed: 14642470]
- Leergaard TB, White NS, De Crespigny A, Bolstad I, D’Arceuil H, Bjaalie JG, Dale AM, 2010. Quantitative histological validation of diffusion MRI fiber orientation distributions in the rat brain. *PLoS One* 5 (1), e8595. [PubMed: 20062822]
- Lehman JF, et al., 2011. Rules ventral prefrontal cortical axons use to reach their targets: implications for diffusion tensor imaging tractography and deep brain stimulation for psychiatric illness. *J. Neurosci.* 31, 10392–10402. [PubMed: 21753016]
- Leuze C, Goubran M, Barakovic M, Aswendt M, Tian Q, Hsueh B, Crow A, Weber EMM, Steinberg GK, Zeineh M, Plowey ED, Daducci A, Innocenti G, Thiran JP, Deisseroth K, McNab JA, 2021. Comparison of diffusion MRI and CLARITY fiber orientation estimates in both gray and white matter regions of human and primate brain. *Neuroimage* 228, 117692. [PubMed: 33385546]
- Lin CP, et al., 2003. Validation of diffusion spectrum magnetic resonance imaging with manganese-enhanced rat optic tracts and ex vivo phantoms. *Neuroimage* 19 (3), 482–495. [PubMed: 12880782]
- Lin CP, et al., 2001. Validation of diffusion tensor magnetic resonance axonal fiber imaging with registered manganese-enhanced optic tracts. *Neuroimage* 14 (5), 1035–1047. [PubMed: 11697935]
- Maffei C, et al., 2020. The IronTract challenge: validation and optimal tractography methods for the HCP diffusion acquisition scheme. *Proc. Intl. Soc. Mag. Res. Med.*
- Maier-Hein KH, et al., 2017. The challenge of mapping the human connectome based on diffusion tractography. *Nat. Commun.* 8 (1), 1349. Available at <http://www.nature.com/articles/s41467-017-01285-x>. [Accessed February 23, 2018]. [PubMed: 29116093]
- Majka P, Wójcik DK, 2016. Possum - a framework for three-dimensional reconstruction of brain images from serial sections. *Neuroinformatics* 14 (3), 265–278. [PubMed: 26687079]
- Martino J, et al., 2011. Cortex-sparing fiber dissection: an improved method for the study of white matter anatomy in the human brain. *J. Anat.* 219 (4), 531–541. Available at <http://www.ncbi.nlm.nih.gov/pubmed/21767263>. [Accessed April 1, 2018]. [PubMed: 21767263]
- Miller KL, et al., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19 (11), 1523–1536. [PubMed: 27643430]
- Mollink J, et al., 2017. Evaluating fibre orientation dispersion in white matter: comparison of diffusion MRI, histology and polarized light imaging. *Neuroimage* 157, 561–574. [PubMed: 28602815]
- Mori S, et al., 1999. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Ann. Neurol.* 45 (2), 265–269. Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9989633. [PubMed: 9989633]
- Moseley ME, et al., 1990. Diffusion-weighted MR imaging of anisotropic water diffusion in cat central nervous system. *Radiology* 176 (2), 439–445. [PubMed: 2367658]
- Neher PF, et al., 2014. Fiberfox: facilitating the creation of realistic white matter software phantoms. *Magn. Reson. Med.* 72 (5), 1460–1470. Available at 10.1002/mrm.25045. [Accessed April 1, 2018]. [PubMed: 24323973]
- Parker GJM, et al., 2002. Initial demonstration of in vivo tracing of axonal projections in the macaque brain and comparison with the human brain using diffusion tensor imaging and fast marching tractography. *Neuroimage* 15 (4), 797–809. Available at <http://www.ncbi.nlm.nih.gov/pubmed/11906221>. [Accessed March 31, 2015]. [PubMed: 11906221]
- Perrin M, et al., 2005. Validation of q-ball imaging with a diffusion fibre-crossing phantom on a clinical scanner. *Philosoph. Trans. R. Soc. London. Series B, Biol. Sci.* 360 (1457), 881–891.
- Poupon C, et al., 2008. New diffusion phantoms dedicated to the study and validation of high-angular-resolution diffusion imaging (HARDI) models. *Magn. Reson. Med.* 60 (6), 1276–1283. [PubMed: 19030160]

- Reisert M, Kellner E, Kiselev VG, 2012. About the geometry of asymmetric fiber orientation distributions. *IEEE Trans. Med. Imaging*31, 1240–1249. [PubMed: 22345527]
- Reuter M, Rosas HD, Fischl B, 2010. Highly accurate inverse consistent registration: a robust approach. *Neuroimage*53 (4), 1181–1196. Available at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2946852&tool=pmcentrez&rendertype=abstract>. [Accessed May 7, 2015]. [PubMed: 20637289]
- Safadi Z, et al., 2018. Functional segmentation of the anterior limb of the internal capsule: linking white matter abnormalities to specific connections. *J. Neurosci.* 2335. –17 Available at <http://www.ncbi.nlm.nih.gov/pubmed/29358360>. [Accessed January 25, 2018].
- Savadjiev P, Campbell JSW, Descoteaux M, Deriche R, Pike GB, Siddiqi K, 2008. Labeling of ambiguous subvoxel fibre bundle configurations in high angular resolution diffusion MRI. *Neuroimage*41 (1), 58–68. [PubMed: 18367409]
- Schilling K, Janve V, Gao Y, Stepniewska I, Landman BA, Anderson AW, 2016. Comparison of 3D orientation distribution functions measured with confocal microscopy and diffusion MRI. *Neuroimage*129, 185–197. [PubMed: 26804781]
- Schilling KG, et al., 2018. Histological validation of diffusion MRI fiber orientation distributions and dispersion. *Neuroimage*165, 200–221. Available at <http://www.sciencedirect.com/science/article/pii/S1053811917308728?via%3Dihub#bib74>. [Accessed December 11, 2017]. [PubMed: 29074279]
- Schilling KG, Gao Y, et al., 2019. Anatomical accuracy of standard-practice tractography algorithms in the motor system - a histological validation in the squirrel monkey brain. *Magn. Reson. Imaging*55, 7–25. Available at <https://www.sciencedirect.com/science/article/pii/S0730725X18302212>. [Accessed October 4, 2018]. [PubMed: 30213755]
- Schilling KG, Nath V, et al., 2019. Limits to anatomical accuracy of diffusion tractography using modern approaches. *Neuroimage*185, 1–11. Available at <https://www.sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S1053811918319888?via%3Dihub>. [Accessed November 19, 2018]. [PubMed: 30317017]
- Schmahmann JD, et al., 2007. Association fibre pathways of the brain: parallel observations from diffusion spectrum imaging and autoradiography. *Brain*130 (3), 630–653. [PubMed: 17293361]
- Seehaus A, Roebroek A, Bastiani M, Fonseca L, Bratzke H, Lori N, Vilanova A, Goebel R, Galuske R, 2015. Histological validation of high-resolution DTI in human post mortem tissue. *Front. Neuroanatomy*9, 98.
- Sotiropoulos SN, et al., 2011. Inference on constant solid angle orientation distribution functions from diffusion-weighted mri.. In: *OHBM, Canada*, p. 609.
- Somerville LH, et al., 2018. The lifespan human connectome project in development: a large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*183, 456–468. [PubMed: 30142446]
- Tang W, et al., 2019. A connectional hub in the rostral anterior cingulate cortex links areas of emotion and cognitive control. *eLife*43761 2019.
- Thomas C, et al., 2014. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. *Proc. Natl. Acad. Sci.* 111 (46), 16574–16579. Available at <http://www.pnas.org/content/111/46/16574.long>. [Accessed November 4, 2014]. [PubMed: 25368179]
- Tobisch A, et al., 2018. Compressed Sensing Diffusion Spectrum Imaging for Accelerated Diffusion Microstructure MRI in Long-Term Population Imaging. *Front. Neurosci.* 12, 650. [PubMed: 30319336]
- Tournier JD, Calamante F, Connelly A, 2012. MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* 22 (1), 53–66.
- Tuch DS, 2004. Q-ball imaging. *Magnet. Res. Med.* 52, 1358–1372.
- Wedeen VJ, et al., 2005. Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. *Magnet. Res. Med.* 54, 1377–1386.
- Whitcher B, et al., 2008. Using the wild bootstrap to quantify uncertainty in diffusion tensor imaging. *Hum. Brain Mapp.* 29 (3), 346–362. [PubMed: 17455199]

- Author Manuscript
- Author Manuscript
- Author Manuscript
- Author Manuscript
- Yamada M, et al., 2008. Diffusion-tensor neuronal fiber tractography and manganese-enhanced MR imaging of primate visual pathway in the common marmoset: preliminary results. *Radiology* 249 (3), 855–864. [PubMed: 19011185]
- Yeh FC, et al., 2013. Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PLoS One* 8 (11). Available at: [dx10.1371/journal.pone.0080713](https://doi.org/10.1371/journal.pone.0080713).
- Yeh FC, Wedeen VJ, Tseng WYI, 2010. Generalized q-sampling imaging. *IEEE Trans. Med. Imaging* 29 (9), 1626–1635. [PubMed: 20304721]
- Yendiki A, et al., 2011. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. *Front. Neuroinformat.* 5, 23. Available at <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3193073&tool=pmcentrez&rendertype=abstract>. [Accessed October 15, 2014].
- Yendiki A, et al., 2020. Towards taking the guesswork (and the errors) out of diffusion tractography. *Proc. Intl. Soc. Mag. Res. Med.*.
- Yendiki A, Aggarwal M, Axer M, Howard AFD, van Cappellen van Walsum A-M, Haber SN (2021). Post mortem mapping of connectional anatomy for the validation of diffusion MRI 10.1101/2021.04.16.440223v1
- Yushkevich PA, et al., 2006. 3D mouse brain reconstruction from histology using a coarse-to-fine approach. In: *Lecture Notes in Computer Science*, vol. 4057, pp. 230–237.

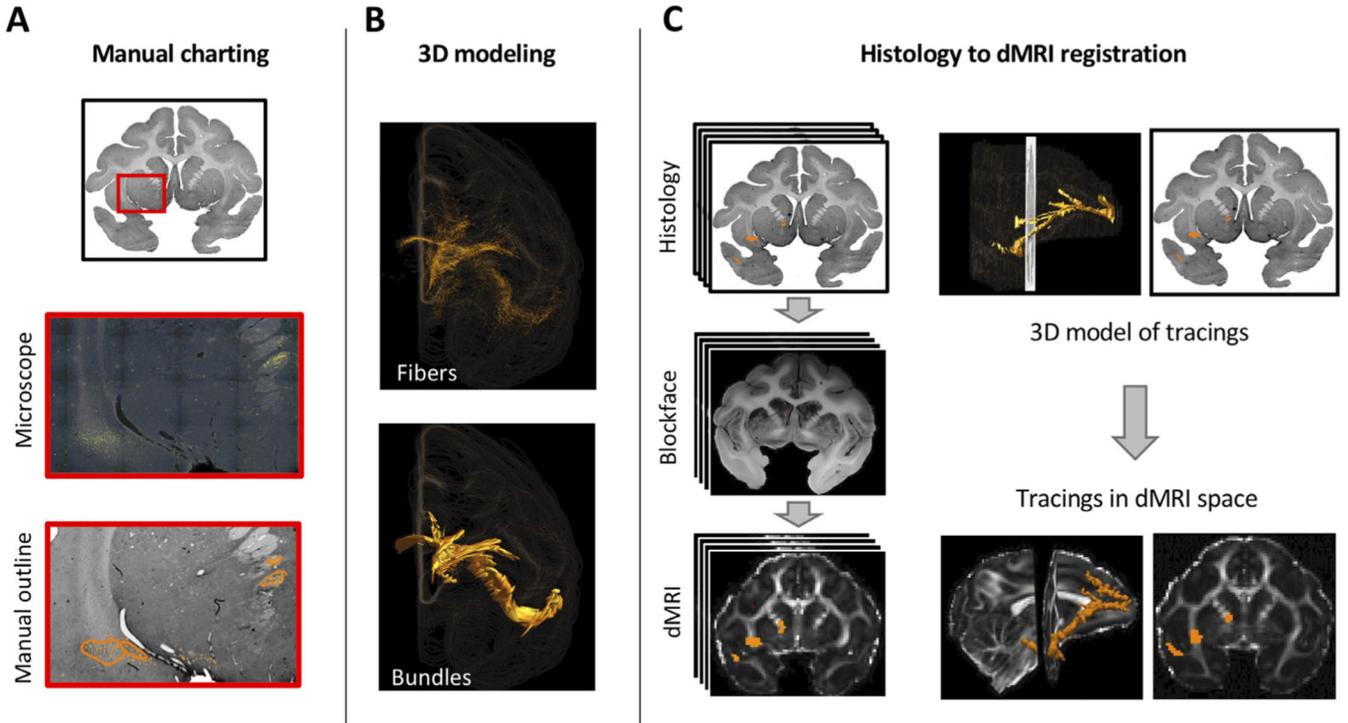


Fig. 1. Comparison of anatomic tracing and dMRI. **A:** Axons labeled by a tracer injection in the frontal pole of a macaque brain were outlined manually under the microscope. **B:** The outlines were combined across sections into a 3D model. *Top:* Individual fiber orientations, which were used to guide the manual outlining of the contours across consecutive sections. *Bottom:* Final 3D rendering of bundles. **C:** Each histological slice was registered to the corresponding blockface photo and the stacked slices were then registered to the dMRI $b = 0$ vol. This transformation was used to transfer the tracing outlines from histology to dMRI space.

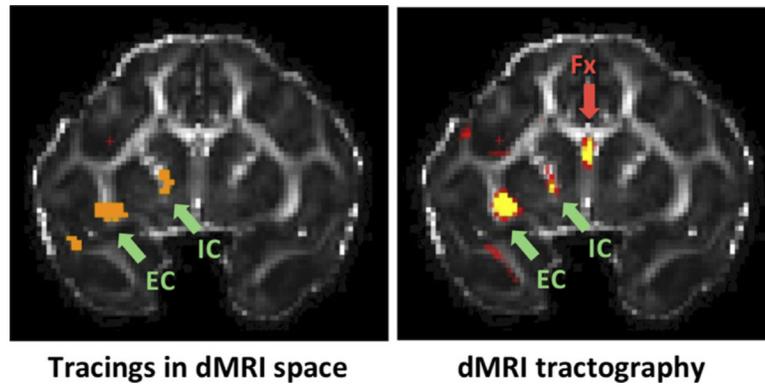


Fig. 2. Definition of true and false positives. Example of voxels labeled by the tracer (left) vs. dMRI tractography (right), displayed on the fractional anisotropy map of the same slice. Voxels labeled by both tracer and tractography, such as in the internal capsule (IC) and the external capsule (EC), are deemed true positives (green arrows). Those labeled by dMRI tractography only, such as in the fornix (Fx) are deemed false positives (red arrow).

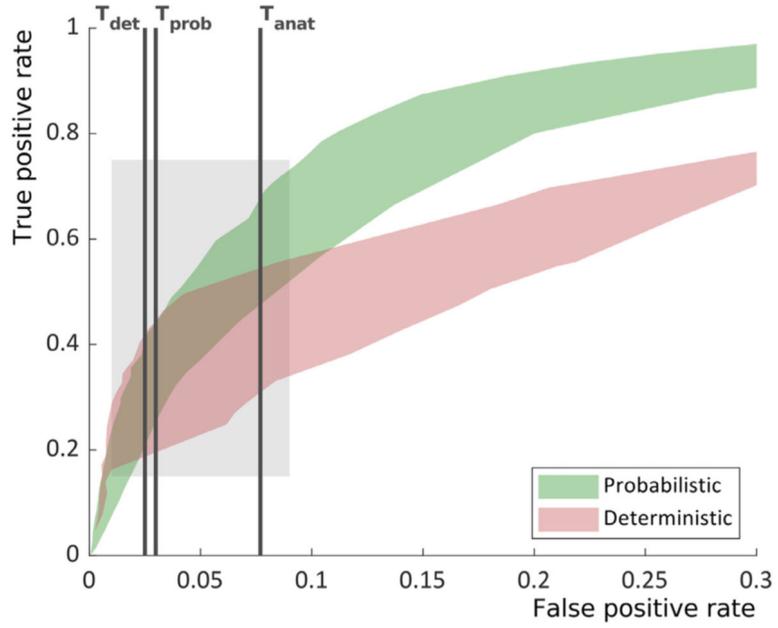


Fig. 3. Summary of ROC curves for dMRI tractography vs. anatomic tracing in the same brain. The figure summarizes the results for all methods. The green shaded area contains the ROC curves of probabilistic tractography with all reconstruction methods and q-space sampling schemes. The red shaded area contains the ROC curves of deterministic tractography with all reconstruction methods and q-space sampling schemes. Vertical lines show the three operating points: T_{det} , T_{prob} , and T_{anat} . The area demarcated by the gray box is magnified in the plots of Fig. 4, where ROC curves are shown separately for each q-space sampling scheme.

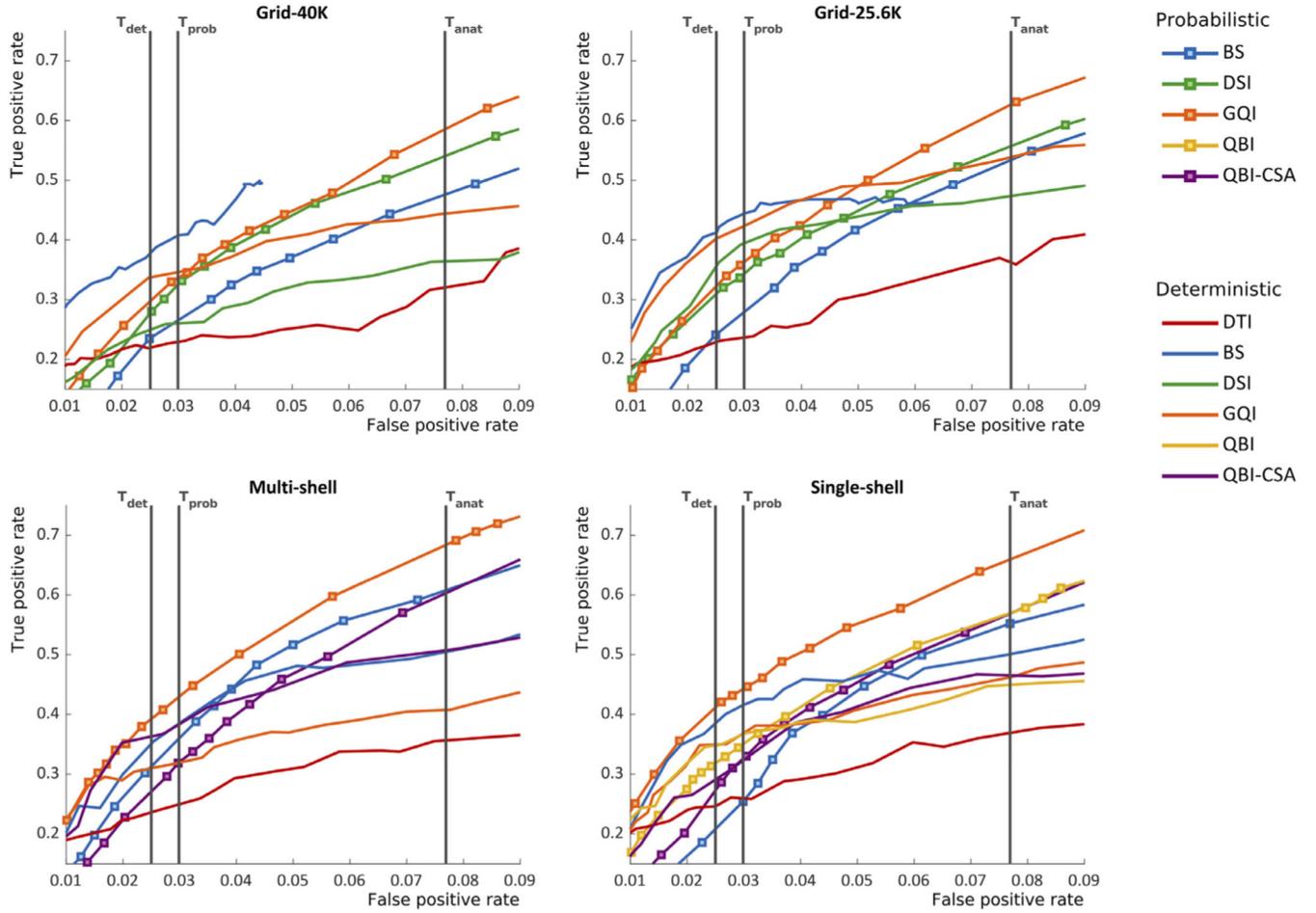


Fig. 4. ROC curves for dMRI tractography vs. anatomic tracing in the same brain. The section of the ROC curves in Fig. 3 that is demarcated by a gray box is shown here in magnification, with the ROC curves grouped by q-space sampling scheme: grid-40 K, grid-25.6 K, multi-shell, and single-shell. Each color represents a different orientation reconstruction method. Curves with and without square markers show results, respectively, from probabilistic and deterministic tractography. Vertical lines show the three operating points: T_{det} , T_{prob} , and T_{anat} .

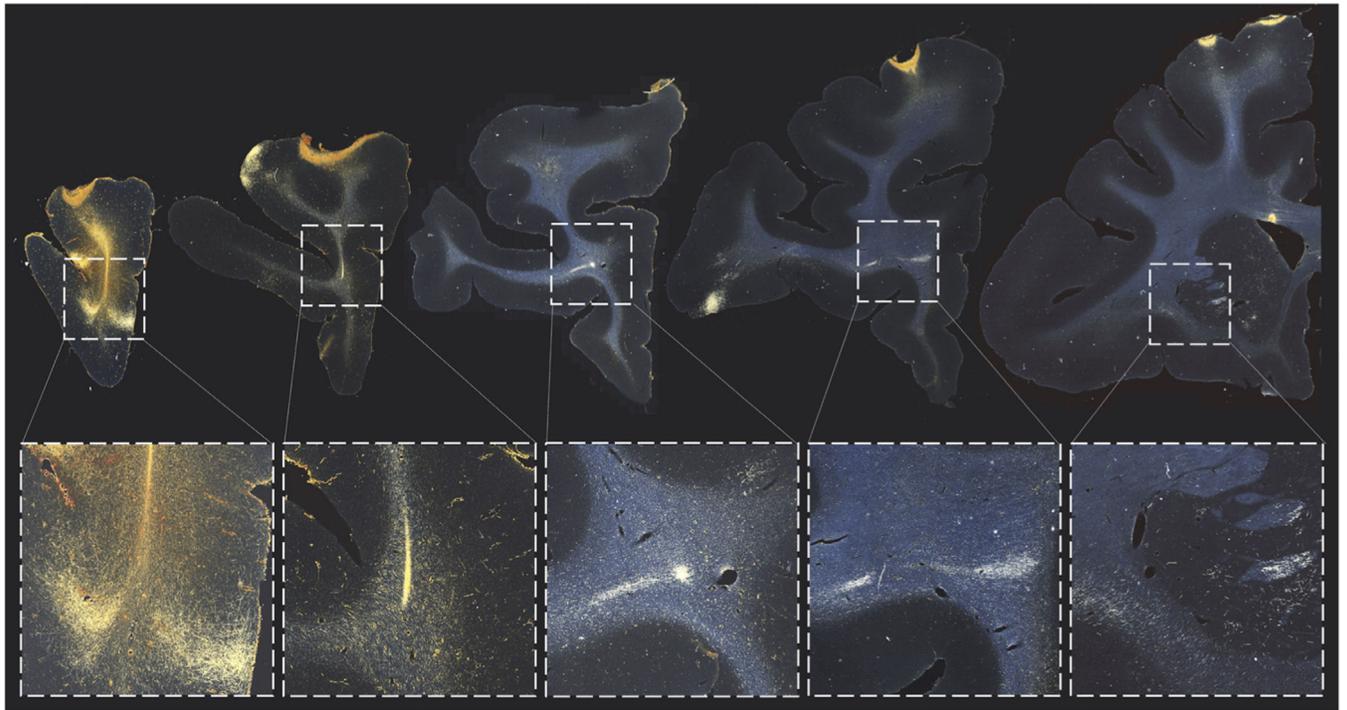


Fig. 5. Injection in the frontopolar cortex. Photomicrographs on the top row show coronal sections, starting near the injection site and moving rostral to caudal. The insets on the bottom row show magnified regions from each of the five sections, with fibers labeled by the tracer. The magnified regions show, from left to right: (i) The stalk and other fibers leaving the injection site. (ii) The stalk as it travels in a caudal and ventral direction. (iii) The stalk branching into two groups of axons. (iv) The two groups of axons, traveling towards the capsules (lateral) and the CC (medial). (v) The former group, after it has further branched into the EC (lateral) and IC (medial).

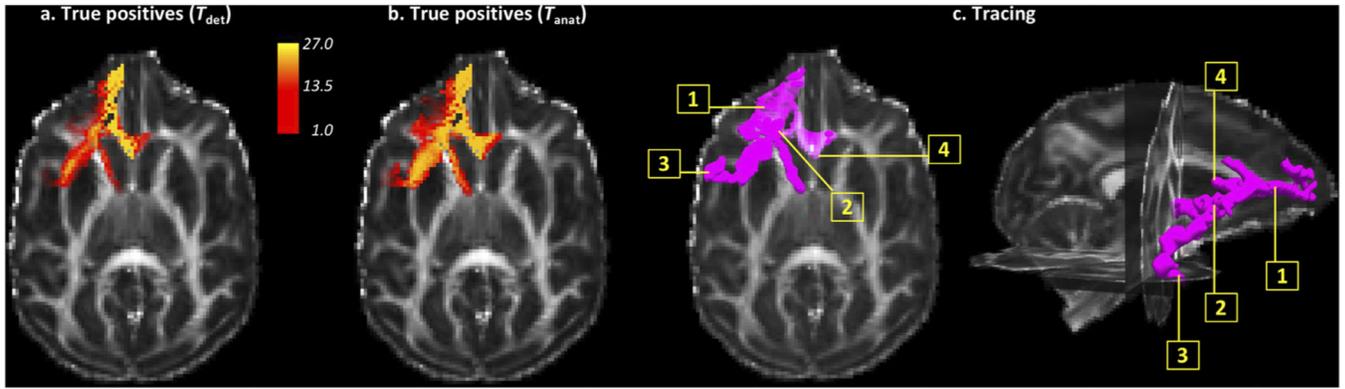


Fig. 6.

Histograms of true positives across tractography methods. (a) Maximum intensity projection through a histogram that shows the number of tractography methods achieving a TP at each voxel, when their FPR is set to that of the default deterministic threshold T_{det} . (b) As above, for the anatomically defined threshold T_{anat} . (c) A 3D isosurface of the tracer injection. Four challenging (fanning, branching, or turning) areas are shown, in both axial and sagittal view: 1. Fanning toward the lateral prefrontal cortex 2. Branching toward the capsules 3. Turning into the uncinate fasciculus 4. Turning along the corpus callosum.

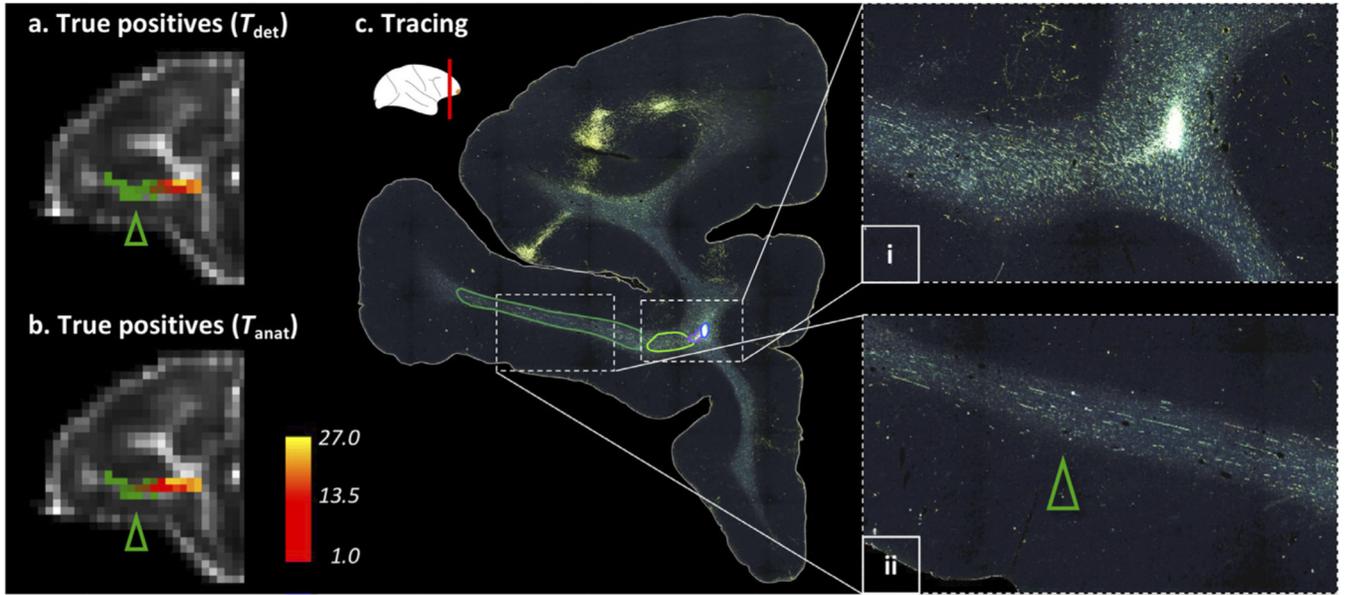


Fig. 7.

Fanning area: lateral PFC. (a) The heat map shows the number of tractography methods achieving a TP at each voxel in a coronal slice, when their FPR is set to that of the default deterministic threshold T_{det} . (b) As above, for the anatomically defined threshold T_{anat} . Both TP heat maps are superimposed on a map of voxels that are known to be connected to the injection site based on the tracer data, including fibers that travel to the lateral PFC (green) and that are missed by tractography. (c) Photomicrograph of the same coronal slice, showing the fibers labeled by the tracer. Two insets show magnifications of: (i) Dense fibers in the main stalk originating from the injection site and (ii) Sparser fibers fanning off the main stalk and toward the lateral PFC, which are missed by tractography.

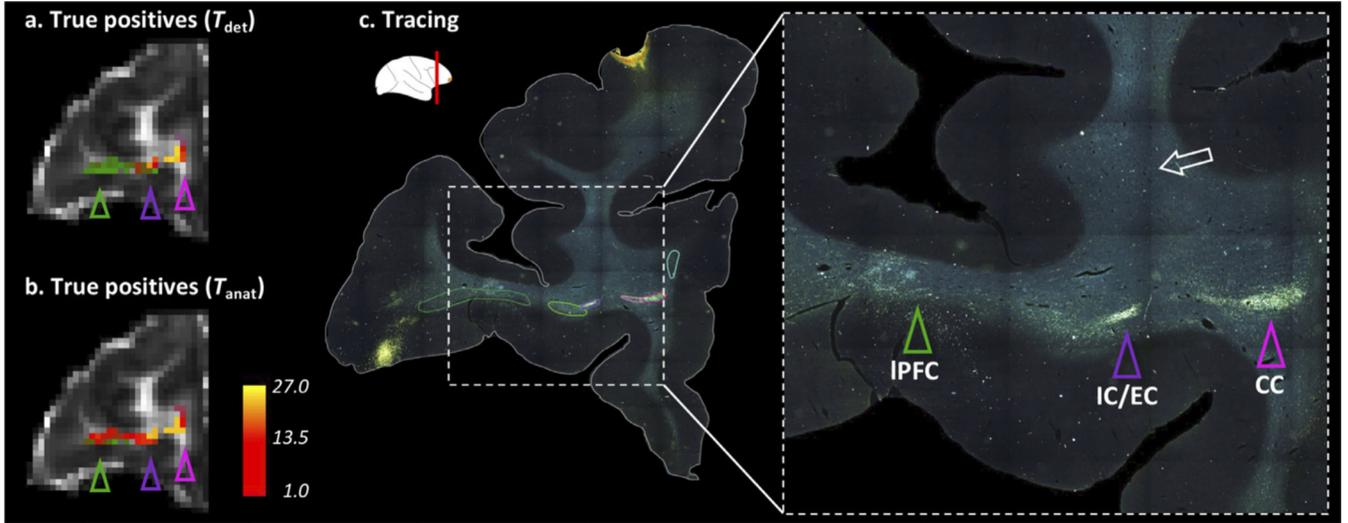


Fig. 8. Branching area: IC/EC. (a) The heat map shows the number of tractography methods achieving a TP at each voxel in a coronal slice, when their FPR is set to that of the default deterministic threshold T_{det} . (b) As above, for the anatomically defined threshold T_{anat} . Both TP heat maps are superimposed on a map of voxels that are known to be connected to the injection site based on the tracer data, including fibers that travel to the lateral PFC (green), IC/EC (purple), and CC (pink). (c) Photomicrograph of the same coronal slice, with outlines indicating different groups of axons labeled by the tracer. The inset shows a magnification of three of these groups of axons, which travel in the lateral PFC, IC/EC, and CC.

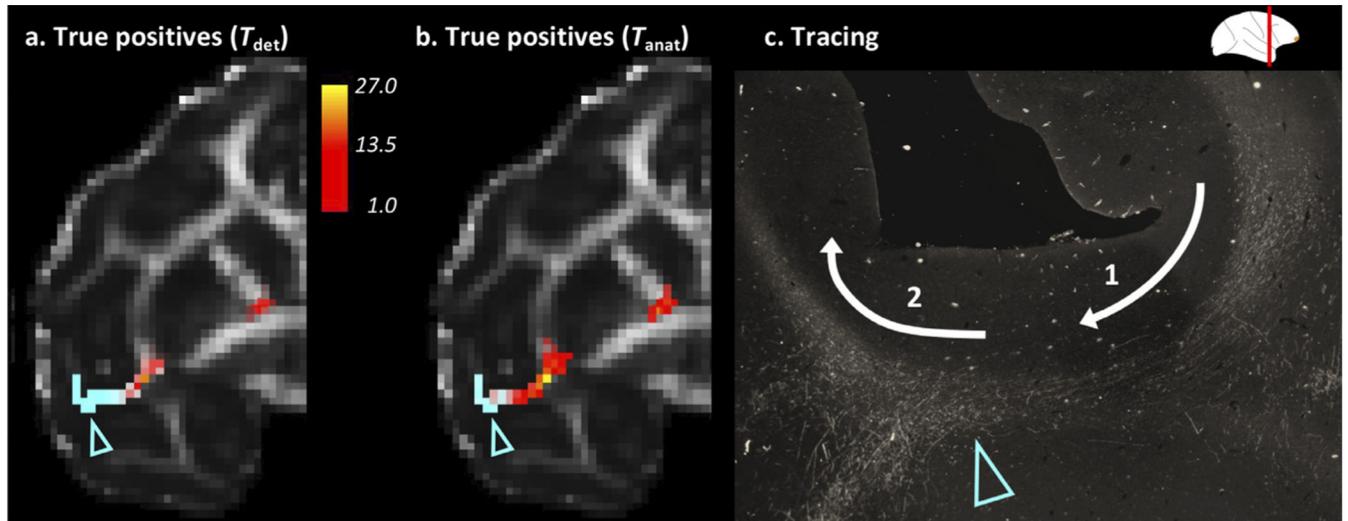


Fig. 9. Turning area: UF. (a) The heat map shows the number of tractography methods achieving a TP at each voxel in a coronal slice, when their FPR is set to that of the default deterministic threshold T_{det} . (b) As above, for the anatomically defined threshold T_{anat} . Both TP heat maps are superimposed on a map of voxels that are known to be connected to the injection site based on the tracer data, including fibers that turn into the UF (blue). (c) Photomicrograph of the same coronal slice, showing the area of the turn. Fibers leaving the EC turn first laterally towards the UF (arrow 1) and then dorsally into the middle longitudinal fasciculus (arrow 2). This double turn cannot be reconstructed by tractography.

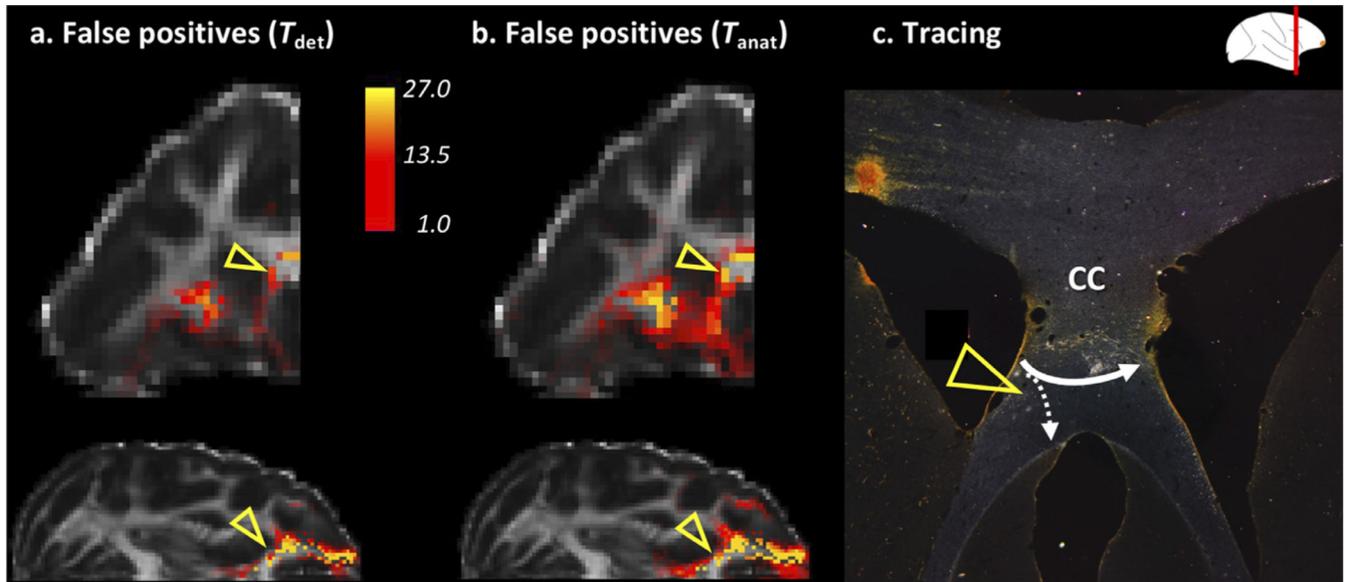
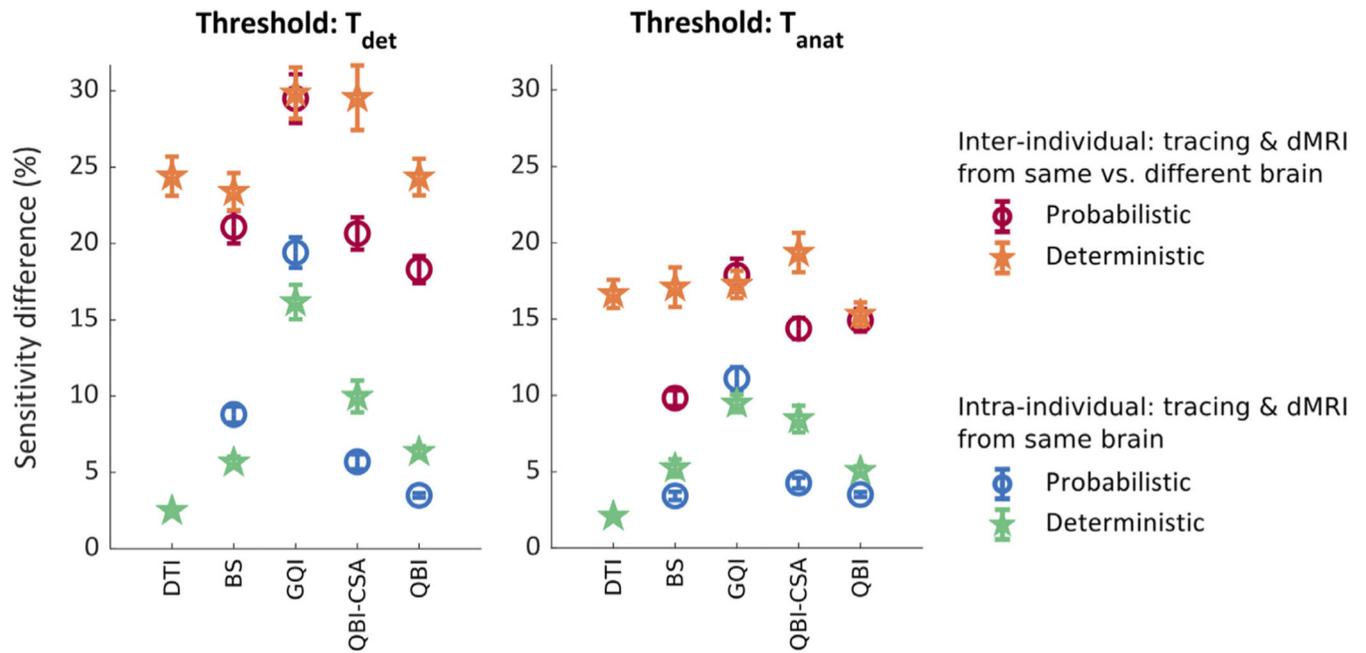


Fig. 10.

Turning area: CC. (a) The heat map shows the number of tractography methods that commit a FP at each voxel in a coronal and an axial slice, when their FPR is set to that of the default deterministic threshold T_{det} . (b) As above, for the anatomically defined threshold T_{anat} . The yellow arrows show an area of consistent FPs in the CC. (c) Photomicrograph of the same coronal slice, showing no fibers traveling in the rostro-caudal direction of the FP tractography streamlines. Instead, fibers from the injection site travel through the genu of the CC and cross to the contralateral hemisphere with high curvature.

**Fig. 11.**

Intra- vs. inter-individual variability. The difference between ROC analyses performed on a dMRI dataset from the same brain as the tracing and a dMRI dataset from a different brain is greater than the difference between ROC analyses performed on two dMRI datasets from the same brain as the tracing. The plots show the difference in TPR at the same FPR level (left: T_{det} ; right: T_{anat}), between different single-shell direction sets from the same brain (blue: probabilistic tractography; green: deterministic tractography) or different brains (red: probabilistic tractography; orange: deterministic tractography). In the latter case, differences were computed between the brain with the injection and each of the other 13 brains. The plots show average differences across brains, with standard error bars.

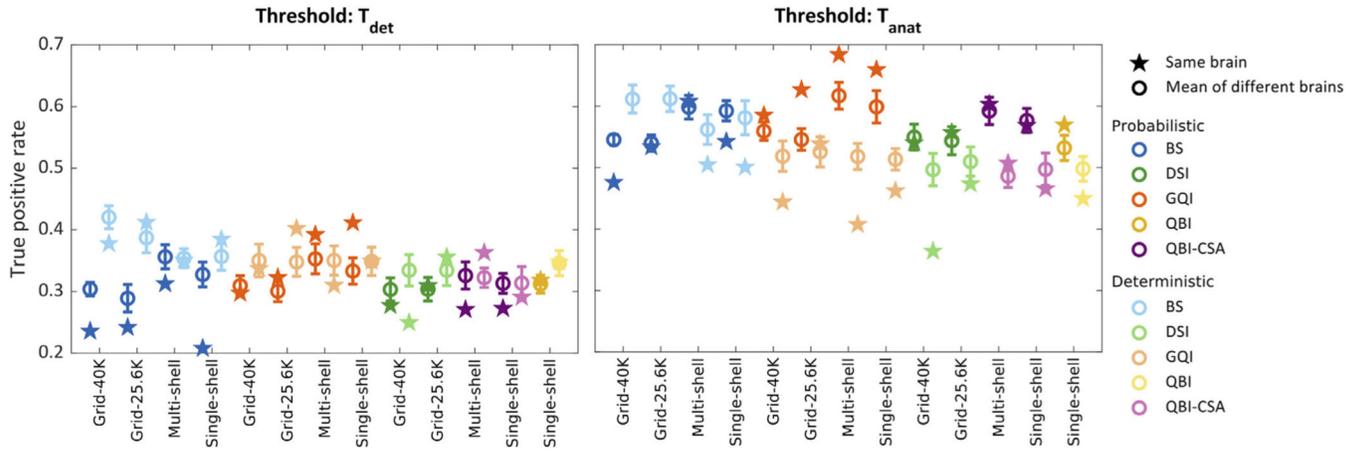


Fig. 12. Sensitivity with respect to tracer data from the same or a different brain. Performing ROC analyses on dMRI data from a set of different brains and averaging the results is not a substitute for performing the ROC analyses directly on dMRI data from the same brain as the tracing. The plots show the TPR at the same FPR level (left: T_{det} ; right: T_{anat}), when the TPs and FPs are identified based on a tracer injection in the same brain (stars) or a different brain (circles). In the latter case, the tracer data are compared to dMRI tractography from 13 different brains, and the plots show average TPR with standard error bars. Each color represents a different dMRI reconstruction method. Lighter/darker shades of the same color denote deterministic/probabilistic tractography.

Table 1

Overview of tractography methods. The table shows all combinations of orientation reconstruction method (row) and q-space sampling scheme (column) that we evaluated. Deterministic and probabilistic tractography are denoted, respectively, by “*d*” and “*p*”.

	Grid-40K	Grid-25.6K	Multi-shell	Single-shell
DTI	<i>d</i>	<i>d</i>	<i>d</i>	<i>d</i>
BS	<i>d,p</i>	<i>d,p</i>	<i>d,p</i>	<i>d,p</i>
GQI	<i>d,p</i>	<i>d,p</i>	<i>d,p</i>	<i>d,p</i>
DSI	<i>d,p</i>	<i>d,p</i>	–	–
QBI-CSA	<i>d,p</i>	–	<i>d,p</i>	<i>d,p</i>
QBI	<i>d,p</i>	–	–	<i>d,p</i>

Table 2

Sensitivity of dMRI tractography methods at the same level of specificity. Results for each method and % difference of probabilistic vs. deterministic methods are shown at the three operating points: T_{det} , T_{prob} , and T_{anat} .

	T_{det}			T_{prob}			T_{anat}					
	Grid 40k	Grid 25.6k	Multi shell	Single shell	Grid 40k	Grid 25.6k	Multi shell	Single shell	Grid 40k	Grid 25.6k	Multi shell	Single shell
Probabilistic	0.24	0.24	0.31	0.21	0.26	0.28	0.36	0.26	0.48	0.53	0.61	0.54
Deterministic	0.38	0.41	0.35	0.38	0.41	0.44	0.38	0.42	-	-	0.5	0.5
% difference	-60.29	-70.52	-12.17	-85.09	-53.72	-58.97	-6.97	-60.81	-	-	16.96	7.69
Probabilistic	0.3	0.32	0.39	0.41	0.34	0.36	0.43	0.44	0.59	0.63	0.68	0.66
Deterministic	0.34	0.4	0.31	0.35	0.35	0.42	0.32	0.37	0.44	0.54	0.41	0.46
% difference	-13.41	-24.68	20.99	15.15	-2.92	-16.67	25.67	16.8	24.14	14.07	40.44	29.94
Probabilistic	0.28	0.31	-	-	0.32	0.34	-	-	0.54	0.56	-	-
Deterministic	0.25	0.36	-	-	0.26	0.39	-	-	0.36	0.47	-	-
% difference	10.08	-14.99	-	-	19.89	-15.01	-	-	32.66	15.01	-	-
Probabilistic	-	-	0.27	0.27	-	-	0.32	0.33	-	-	0.6	0.57
Deterministic	-	-	0.36	0.29	-	-	0.38	0.32	-	-	0.51	0.47
% difference	-	-	-34.03	-6.65	-	-	-19.89	0.88	-	-	15.89	18.19
Probabilistic	-	-	-	0.32	-	-	-	0.35	-	-	-	0.57
Deterministic	-	-	-	0.35	-	-	-	0.37	-	-	-	0.45
% difference	-	-	-	-9.29	-	-	-	-4.79	-	-	-	21.08