



Published in final edited form as:

Neuroimage. 2017 November 01; 161: 149–170. doi:10.1016/j.neuroimage.2017.08.047.

Harmonization of multi-site diffusion tensor imaging data

Jean-Philippe Fortin¹, Drew Parker², Birkan Tunç², Takanori Watanabe², Mark A. Elliott³, Kosha Ruparel⁴, David R. Roalf⁴, Theodore D. Satterthwaite⁴, Ruben C. Gur^{3,4}, Raquel E. Gur^{3,4}, Robert T. Schultz⁵, Ragini Verma^{2,†}, and Russell T. Shinohara^{1,†,*}

¹Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania

²Center for Biomedical Image Computing and Analytics, Department of Radiology, Perelman School of Medicine, University of Pennsylvania

³Department of Radiology, Perelman School of Medicine, University of Pennsylvania

⁴Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania

⁵Center for Autism Research, The Children's Hospital of Philadelphia

Abstract

Diffusion tensor imaging (DTI) is a well-established magnetic resonance imaging (MRI) technique used for studying microstructural changes in the white matter. As with many other imaging modalities, DTI images suffer from technical between-scanner variation that hinders comparisons of images across imaging sites, scanners and over time. Using fractional anisotropy (FA) and mean diffusivity (MD) maps of 205 healthy participants acquired on two different scanners, we show that the DTI measurements are highly site-specific, highlighting the need of correcting for site effects before performing downstream statistical analyses. We first show evidence that combining DTI data from multiple sites, without harmonization, may be counter-productive and negatively impacts the inference. Then, we propose and compare several harmonization approaches for DTI data, and show that ComBat, a popular batch-effect correction tool used in genomics, performs best at modeling and removing the unwanted inter-site variability in FA and MD maps. Using age as a biological phenotype of interest, we show that ComBat both preserves biological variability and removes the unwanted variation introduced by site. Finally, we assess the different harmonization methods in the presence of different levels of confounding between site and age, in addition to test robustness to small sample size studies.

*To whom correspondence should be addressed. rshi@mail.med.upenn.edu.

†Equal contribution

Competing interests

The authors declare that they have no competing interests.

Authors contributions

JPF developed the methodology and analyzed the data. DP, BT and TW processed the data. ME, KR, DR, TS, RCG, REG and RTSc recruited the participants and acquired the data. JPF and RTSh wrote the manuscript. RTSh and RV supervised the work. All authors read and approved the final manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

DTI; Diffusion; Harmonization; Multi-Site; ComBat; Inter-scanner

1 Introduction

Diffusion tensor imaging (DTI) is a well-established magnetic resonance imaging (MRI) technique for studying the white matter (WM) organization and tissue characteristics of the brain. Diffusion tensor imaging has been used extensively to study both brain development and pathology; see Alexander et al. [2007] for a review of DTI and several of its applications. In studies assessing white matter tissue characteristics, two commonly reported complementary scalar maps are the mean diffusivity (MD), which assesses the degree to which water diffuses at each location, and fractional anisotropy (FA), which measures the coherence of this diffusion in one particular direction. Together, MD and FA provide complementary description of white matter microstructure.

With the increasing number of publicly available neuroimaging databases, a crucial goal is to combine large-scale imaging studies to increase the power of statistical analyses to test common biological hypothesis. For instance, for life-span studies, combining data across sites and age ranges is essential for obtaining the necessary number of participants of each age. The success of combining multi-site imaging data depends critically on the comparability of the images across sites. As with other imaging modalities, DTI images are subject to technical variability across scans, including heterogeneity in the imaging protocol, variations in the scanning parameters and differences in the scanner manufacturers [Zhu et al., 2009, 2011]. Among others, the reliability of FA and MD maps have been shown to be affected by angular and spatial resolution [Zhan et al., 2010, Alexander et al., 2001, Kim et al., 2006], the number of diffusion weighting directions [Giannelli et al., 2009], the number of gradient sampling orientations [Jones, 2004], the number of b-values [Correia et al., 2009], and the b-values themselves.

In the design of multi-site studies, defining a standardized DTI protocol is a first step towards reducing inter-scanner variability. However, even in the presence of a standardized protocol, systematic differences between scanner manufacturers, field strength and other scanner characteristics will systematically affect the DTI images and induce inter-scanner variation. Image-based meta analysis (IBMA) techniques, reviewed in Salimi-Khorshidi et al. [2009], are common methods for combining results from multi-site studies with the goal of testing a statistical hypothesis. IBMA methods circumvent the need of harmonizing images across sites by performing site-specific statistical analyses and combining results afterwards. Fisher's p-value combining method and Stouffer's z-transformation test, applied to z or t-maps, are two common IBMA techniques. Fixed-effect models based on (possibly) normalized images, and mixed-effect models to model the inter- and intra- site variability, are other common techniques for the analysis of multi-site data. Indeed, meta-analysis methods have shown great promise for studies with a large number of participants at each site. For instance, the ENIGMA-DTI working group has been successfully using and

validating meta-analysis techniques on such multi-site DTI data [Jahanshad et al., 2013, Kochunov et al., 2014].

Meta-analysis techniques have several limitations, however. First, study-specific samples might not be sufficient to estimate the true biological variability in the population [Mirzaalian et al., 2016]. As described by De Wit et al. [2014], adjusting for variability at the participant level is problematic in meta-analyses, since only group-level demographic and clinical information is available. Another limitation is that for a multi-site study, computing site-specific summary statistics will be affected by unbalanced data. For instance, the calculation of a variance using unbalanced datasets is highly affected by the ratio cases/controls in the sample [Linn et al., 2016b]. Another limitation, for imaging studies with small sample sizes, the parameters of the z-score transformations cannot be robustly estimated, yielding suboptimal statistical inferences.

Mega-analyses, in which the imaging data are combined before performing statistical inferences, have the potential to increase power compared to meta-analyses [De Wit et al., 2014]. In addition, pooling imaging data across studies has the benefit of enriching the clinical picture of the sample by increasing the variability in symptom profiles [Turner, 2014] and demographic variables. This is particularly important for age-span studies. However, pooling data across studies may increase the heterogeneity of the imaging measurements by introducing undesirable variability caused by differences in scanner protocols. Harmonization of the pooled data is therefore necessary to ensure the success of mega-analyses. The DTI harmonization technique proposed in Mirzaalian et al. [2016] is a first step towards that direction. The method is based on rotation invariant spherical harmonics (RISH) and combines the unprocessed DTI images across scanners. Unfortunately, a major drawback of the method is that it requires DTI data to have similar acquisition parameters across sites, an assumption often infeasible in multi-site observational analyses.

In this work, we adapted and compared several statistical approaches for the harmonization of DTI studies that were previously developed for other data types: Functional normalization [Fortin et al., 2014], RAVEL [Fortin et al., 2016a], Surrogate variable analysis (SVA) [Leek and Storey, 2007] and ComBat [Johnson et al., 2007], a popular batch adjustment method developed for genomics data. We also include a simple method that globally rescales the data for each site using a z-score transformation map common to all features, which we refer to as “global scaling”. For the evaluation of the different harmonization techniques, we use DTI data acquired as a part of two large imaging studies ([Satterthwaite et al., 2014] and [Ghanbari et al., 2014]) with images acquired on different scanners, using different imaging protocols. The participants are teenagers, and were matched across studies for age, gender, ethnicity, and handedness.

We first analyze site-related differences in the FA, MD, radial diffusivity (RD) and axial diffusivity (AD) measurements, and show evidence of significant site effects that differ across the brain. This motivates the need for a harmonization technique that is sensitive to region-specific scanner effects. Then, we harmonize the data with several proposed harmonizations, and evaluate their performance using a comprehensive evaluation

framework. We show that the ComBat is the most effective harmonization techniques as it removes unwanted variation induced by site, while preserving between-subject biological variability. ComBat is a promising harmonization technique for other imaging modalities since it does not make assumptions about the origin of the site effects.

2 Methods

2.1 Data

We consider two DTI studies from two different scanners. To investigate the effect of scanner variations on the DTI measurements, we matched the participants for age, gender, ethnicity and handedness, resulting in 105 participants retained in each study for further analysis. The characteristics of each dataset are described below.

Dataset 1 (Site 1): PNC dataset: We selected a subset of the Philadelphia Neurodevelopmental Cohort (PNC) [Satterthwaite et al., 2014], and included 105 healthy participants from 8 to 19 years old. 83 of the participants were males (22 females), and 75 participants were white (30 non-white). The DTI data were acquired on a 3T Siemens TIM Trio whole-body scanner, using a 32 channel head coil and a twice-refocused spin-echo (TRSE) single-shot EPI sequence with the following parameters: TR = 8100 ms and TE = 82 ms, b-value of 1000 s/mm², 7 $b = 0$ images and 64 gradient directions. The images were acquired at $1.875 \times 1.875 \times 2$ mm resolution. During the same session, structural T1-weighted (T1-w) MP-RAGE images were also acquired with parameters TR = 1810 ms, TE = 3.5 ms, TI = 1100 ms and FA = 9°, at $0.9375 \times 0.9375 \times 1$ mm resolution.

Dataset 2 (Site 2): ASD dataset: The dataset contains 105 typically developing controls (TDC) from a study focusing on autism spectrum disorder (ASD) [Ghanbari et al., 2014]. 83 of the participants were males (22 females), and 79 participants were white (26 non-white). The age of the participants ranges from 8 to 18 years old. The DTI data were acquired on a Siemens 3T Verio scanner, using a 32 channel head coil and a single shot spin-echo planar sequence with the following parameters: TR = 11,000 ms and TE = 76 ms, b-value of 1000 s/mm², 1 $b = 0$ image and 30 gradient directions. The images were acquired at 2mm isotropic resolution. Structural T1-w MP-RAGE images were also acquired with parameters TR = 1900 ms, TE = 2.54 ms, TI = 900 ms and FA = 9° at resolution $0.8 \text{ mm} \times 0.8 \text{ mm} \times 0.9 \text{ mm}$.

For benchmarking the different harmonization procedures, we use two additional subsets of the PNC database, with participants who differ from Dataset 1:

Independent Dataset 1: The dataset contains 292 additional healthy participants from the PNC with the same age range as Dataset 1 and Dataset 2 (8 to 18 years old).

Independent Dataset 2: The dataset contains 105 additional healthy participants from the PNC with an age range of 14 to 22 years old.

2.2 Image processing

Quality control on diffusion weighted images was performed manually. For each DWI volume, we removed weighted gradient images exhibiting signal dropout, likely caused by subject motion and pulsating flow, ghosting artifacts and image stripping. DWI volumes with more than 10% of the weighted images removed, or with a compromised b0 image, were excluded. DWI data were denoised using a joint anisotropic LMMSE filter for Rician noise [Tristán-Vega and Aja-Fernández, 2010] implemented in Slicer (v3.4). The b0 was extracted and skull-stripped using FSL's BET tool [Smith, 2002] (v4.1.5), and the DTI model was fit within the brain mask using an unweighted linear least-squares method. Subsequently, the FA, MD, AD and RD maps were calculated from the resultant tensor image using the python package `dipy` [Garyfallidis et al., 2014] (v0.10.1). The four scalar images were co-registered to the T1-w image using FSL's `flirt` tool [Jenkinson and Smith, 2001, Jenkinson et al., 2002] (v4.1.5) and then non-linearly registered to the Eve template using DRAMMS [Ou et al., 2011] (v1.4.1). We note that the two resulting registration warping transformations were concatenated and applied as a single warp to the scalar DTI maps. A 3-tissue class T1-w segmentation was performed using FSL's FAST tool [Zhang et al., 2001] (v4.1.5) to obtain GM, WM and CSF labels.

2.3 Harmonization methods

We propose to use and adapt five statistical harmonization techniques for DTI data: global scaling, functional normalization [Fortin et al., 2014], RAVEL [Fortin et al., 2016a], Surrogate Variable Analysis (SVA) [Leek and Storey, 2007, 2008] and ComBat [Johnson et al., 2007]. We refer to the absence of harmonization as “raw” data. We now describe the five different methods below with their implementation to the current datasets. For brevity, the different methods are presented in the context of FA intensities, but are similar for MD intensities and other modalities. We use the notation y_{ijv} to denote the FA measure at site i , for scan j and voxel v .

2.3.1 Global scaling—The global scaling (GS) approach is a model that assumes that the effect of each site on the DTI intensities can be summarized into a pair of a global shift and scale parameters $(\theta_{i,location}, \theta_{i,scale})$. More specifically, taking the average intensity map across all sites as a virtual reference site, the location parameter $\theta_{i,location}$ and the scale parameter $\theta_{i,scale}$ for site i can be obtained by fitting the linear model

$$\bar{Y}_i = \theta_{i,location} + \theta_{i,scale} \bar{Y} + E_i, \quad (1)$$

where \bar{Y}_i is the $p \times 1$ average vector of FA intensities for site i , \bar{Y} is the $p \times 1$ global average vector of FA intensities across sites and E_i is a vector of residuals assumed to have mean 0 and variance σ^2 . Estimates $\hat{\theta}_{i,location}$ and $\hat{\theta}_{i,scale}$ can be obtained by ordinary least squares (OLS). To remove the effect of site i on the data, we set the GS-harmonized FA intensity for voxel v and for scan j to be

$$y_{ijv}^{\text{GS}} = \frac{y_{ijv} - \hat{\theta}_{i, \text{location}}}{\hat{\theta}_{i, \text{scale}}}.$$

A more flexible model would be to fit a LOESS or LOWESS curve [Cleveland, 1979, 1981] at each site separately to allow for nonlinearity. This idea was previously used in the so-called loess normalization [Bolstad et al., 2003].

2.3.2 Functional normalization—We apply the functional normalization algorithm, described in Fortin et al. [2014] and later refined in Klein et al. [2015]. Unlike quantile normalization [Bolstad et al., 2003], which forces the histograms to be all the same across subjects, functional normalization only removes variation in the histograms that can be explained by a covariate. It has been successfully used to normalize cancer data and to normalize data from different genomic array technologies [Fortin et al., 2016b]. For multi-site DTI studies, we use site as a covariate. Briefly, the algorithm removes the site effect in the marginal distribution of the FA intensities by modeling the variation in the quantile functions as a function of site. After correction, the marginal densities of the FA intensities are more similar across sites.

2.3.3 RAVEL—The RAVEL algorithm described in Fortin et al. [2016a] attempts to estimate a voxel-specific unwanted variation term by using a control region of the brain to estimate latent factors of unwanted variation common to all voxels. It is an extension of a previous intensity normalization, called White Stripe [Shinohara et al., 2014], developed to normalize white matter intensities in structural MRI. Similar to the control region used in Fortin et al. [2016a], we use voxels labelled as CSF as a control region. Theoretically, the FA values in CSF should be near 0 and similar across participants. In practice, because of the image reconstruction process and differences in protocols and scanning parameters, substantial fluctuations in the FA measurements in CSF exist. For instance, we observe in our dataset an average FA difference of 0.06 between the two sites. Assuming the fluctuations in the FA measurements in CSF are technical in nature, RAVEL uses FA values in CSF as a proxy for technical variability. In Figure B.1a, we show a strong correlation between average FA in the WM and average FA in the CSF. The FA values in the CSF can therefore act as surrogates for site effects in the WM.

Similar to RUV [Gagnon-Bartsch and Speed, 2012], we use singular value decomposition (SVD) to obtain k latent factors of unwanted variation, denoted $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k$, estimated from the CSF control voxels. Using cross-validation, we retain only the first latent factor for further analysis. At each voxel v in the WM, we fit the following linear regression model

$$y_{ijv} = \alpha_v + \psi_v w_{1ij} + \varepsilon_{ijv}$$

to obtain the voxel-specific RAVEL coefficients $\hat{\psi}_v$ (shown in Figure B.1b in template space). We set the RAVEL-harmonized intensity to be $y_{ijv}^{\text{RAVEL}} = y_{ijv} - \hat{\psi}_v w_{1ij}$. The results for MD maps are shown in Figure B.1c–d.

2.3.4 SVA—The SVA algorithm estimates latent factors of unwanted variation, called surrogate variables, that are not associated with the biological covariates of interest. It is particularly useful when the site variable is not known, or when there exists residual unwanted variation after the removal of site effects. We used the reference implementation of SVA in the sva package [Leek et al., 2012], and surrogate variables were estimated using the iteratively re-weighted SVA algorithm [Leek and Storey, 2008]. We provided age and gender as covariates of interest to include in the regression models. The algorithm returns s surrogate variables $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_s$, where s is estimated internally by the algorithm. Similar to RAVEL, we fit at each voxel v in the WM the following linear regression model

$$y_{ijv} = \alpha_v + \sum_{l=1}^s \phi_{lv} z_{lij} + \varepsilon_{ijv}$$

to obtain estimates $\hat{\phi}_{1v}, \hat{\phi}_{2v}, \dots, \hat{\phi}_{sv}$. We set the SVA-harmonized intensity to be

$$y_{ijv}^{\text{SVA}} = y_{ijv} - \sum_{l=1}^s \hat{\phi}_{lv} z_{lij}.$$

2.3.5 ComBat—The ComBat model was introduced in the context of gene expression analysis by Johnson et al. [2007] as an improvement of location/scale models [Parmigiani et al., 2003] for studies with small sample size. Here, we reformulate the ComBat model in the context of DTI images. We assume that the data come from m imaging sites, containing each n_i scans for $i = 1, 2, \dots, m$. For voxel $v = 1, 2, \dots, p$, let y_{ijv} represent the FA measure for the scan j at site i . After some standardization discussed in Johnson et al. [2007], ComBat posits the following location and scale (L/S) adjustment model:

$$y_{ijv} = \alpha_v + \mathbf{X}_{ij} \boldsymbol{\beta}_v + \gamma_{iv} + \delta_{iv} \varepsilon_{ijv}, \quad (2)$$

where α_v is the overall FA measure for voxel v , \mathbf{X} is a design matrix for the covariates of interest (e.g. gender, age), and $\boldsymbol{\beta}_v$ is the voxel-specific vector of regression coefficients corresponding to \mathbf{X} . We further assume that the error terms ε_{ijv} follow a normal distribution with mean zero and variance σ_v^2 . The terms γ_{iv} and δ_{iv} represent the additive and multiplicative site effects of site i for voxel v , respectively.

ComBat uses an empirical Bayes (EB) framework to improve the variance of the parameter estimates $\hat{\gamma}_{iv}$ and $\hat{\delta}_{iv}$. It estimates an empirical statistical distribution for each of those parameters by assuming that all voxels share the same common distribution. In that sense, information from all voxels is used to inform the statistical properties of the site effects. More specifically, the site-effect parameters are assumed to have the parametric prior distributions:

$$\gamma_{iv} \sim N(\gamma_i, \tau_i^2) \quad \text{and} \quad \delta_{iv}^2 \sim \text{Inverse Gamma}(\lambda_i, \theta_i). \quad (3)$$

The hyperparameters γ_i , τ_i^2 , λ_i , θ_i are estimated empirically from the data as described in Johnson et al. [2007]. In Figure 1, we present the distributions of the voxel-wise estimates γ_{iv} and δ_{iv}^2 for each site (dotted lines) together with the estimated prior distributions (solid lines); the estimated prior distributions fit the data well. We note that the `sva` package also offers the option to posit non-parametric priors for more flexibility, at the cost of increasing computational time. As described in Johnson et al. [2007], the ComBat estimates γ_{iv}^* and δ_{iv}^* of the site effect parameters are computed using conditional posterior means, and are shown in Figure 1c in template space.

The final ComBat-harmonized FA values are defined as

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \hat{\alpha}_v - \mathbf{X}_{ij}\hat{\beta}_v - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + \mathbf{X}_{ij}\hat{\beta}_v$$

In Appendix A.1, we present a straightforward generalization of the ComBat model for nonlinear effects of the biological covariates on the imaging measurements using cubic splines. This is motivated by the problem of combining multi-site data for the study of life-span trajectories, for which it is common to observe nonlinear associations between age and imaging measurements [Westlye et al., 2009, Lebel et al., 2012].

2.4 Evaluation framework

We consider a harmonization method to be successful if: (1) it removes the unwanted variation induced by site, scanner or differences in imaging protocols; (2) it preserves between-subject biological variability. Both conditions must be simultaneously tested on the same set of images; it is pointless to remove the noise associated with site if we cannot concurrently maintain the biological variation.

To evaluate (1), we calculate two-sample t-tests on the DTI intensities, testing for a difference between Site 1 and Site 2 measurements. We perform the analysis both at the voxel and ROI level. A harmonization technique that successfully removes site effect will result in non-significant tests, after possibly correcting for multiple comparisons. We base our evaluation of (2) on the replicability and validity of voxels associated with biological variation, using age as the biological factor of interest. Replicability is defined as the chance that an independent experiment will produce a similar set of results [Leek and Peng, 2015], and is a strong indication that a set of results is biologically meaningful. Associations with age are measured using usual Wald t-statistics from linear regression. We test the replicability of the voxels associated with age using a discovery-validation scheme.

In the discovery-validation scheme, we consider the harmonized dataset (Site 1 + Site2 + Harmonization) as a discovery cohort. For the validation cohort, we consider an independent dataset with unrelated participants. In this paper, two independent datasets are considered (see Data section). We then perform a mass-univariate analysis that estimates a t-statistic at each voxel testing for association with age, for the discovery and validation cohorts separately. We denote the two vectors of t-statistics t_{dis} and t_{val} respectively.

In the case of a successful harmonization, the vector t_{dis} should be more similar to the vector t_{val} . While one could use the usual Pearson correlation coefficient to test for the replicability of the two vectors, this has the drawback of considering all voxels equally (both signal and noise voxels), and therefore the measure of replicability is not restrained to the voxels of interest. Because we wish to test for the replicability of the signal voxels only (voxels associated with age), we instead use concordance at the top (CAT) curves [Irizarry et al., 2005]. The CAT curves estimate the overlap between the top k t-statistics, which are the voxels most likely associated with age, for all possible values of k . A CAT curve closer to 1 indicates better overlap between the two lists of t-statistics. We summarize the discovery-validation scheme for replicability in Figure B.2.

2.5 Creation of silver-standards

To further evaluate the performance of the different harmonization methods, we create two sets of silver-standards: a silver-standard for voxels that are truly associated with age (signal silver-standard), and one for voxels not associated with age (null silver-standard).

2.5.1 Creation of a silver-standard for voxels associated with age—Many regions in the WM have been previously demonstrated to show an increase of FA values through adolescence, accompanied by decreasing of MD values [Tamnes et al., 2010, Bava et al., 2010, Lebel and Beaulieu, 2011]. Because some of the reported regions are specific to FA only, or specific to MD only, we estimate a reference set of voxels that substantially change with age for FA, and an additional set for substantial changes in the MD maps, for each site separately. Because our reference sets are estimated within site, they are free of site effects and should represent the best silver-standards for voxels associated with age: we refer to those sets as *signal silver-standards*. To estimate the signal silver-standard for FA (and similarly for MD), we use the following meta-analytic approach: for each site separately, at each voxel in the WM, we apply a linear regression model to obtain a t-statistic measuring the association of FA with age. For each site, we define the site-specific signal silver-standard to be the $k = 5000$ voxels with the highest t-statistics in magnitude. We define the signal silver-standard to be the intersection of the two site-specific signal silver standards. This ensures that the resulting voxels are not only voxels highly associated with age within a study, but are also replicated across the two sites.

For the FA values, this resulted in 2265 voxels for the signal silver-standard set. Among those voxels, 21.3% are located in the thalamic region, 17.1% are located in the anterior limb of the internal capsule (left and right) 14.8% in the posterior limb of the internal capsule (left and right), 10.8% in the midbrain, 9.7% in the cerebral peduncle and 4.9% in the globus pallidus. These results are highly consistent with the changes reported in literature for the same age group [Schmithorst et al., 2002, Barnea-Goraly et al., 2005, Ashtari et al., 2007, Bava et al., 2010, Giorgio et al., 2010]. For the MD values, we obtained a signal silver-standard set of 1932 voxels. 30.4% of these voxels are located in the superior corona radiata, 15.0% are located in the superior frontal lobe, 10.1% are located in the precentral region, 9.4% are located in the superior longitudinal fasciculus, 7.9% are located in the middle frontal region and 6.4% are located in the thalamic region, which is consistent

with regions previously reported in the literature [Bava et al., 2010, Tamnes et al., 2010, Krogsrud et al., 2016].

2.5.2 Creation of a silver-standard for voxels not associated by age—In addition to a signal silver-standard for voxels associated with age, we created silver-standard sets for voxels unaffected by age, for both FA and MD maps, that we refer to as *null silver-standards*. Our approach is similar to that of signal silver-standards. For each site separately, at each voxel in the WM, we apply a linear regression model to obtain a t-statistic measuring the association of FA with age. For each site, we define the site-specific silver-standard for null voxels to be the $k = 5000$ voxels with the lowest t-statistics in magnitude (close to 0). We define the silver-standard to be the intersection of the two site-specific silver standards. This ensures that the resulting voxels are voxels with least association with age within a study, and are also replicated across the two sites.

For the FA values, we obtained a silver-standard set of 405 voxels. We note that this replication rate (8.1%) is much more lower than the replication rate for the signal silver-standard set (45.3%). This is not surprising; strong signal voxels are more likely to replicate than noise voxels. The top regions represented in the null silver-standard are the middle frontal lobe (8.6%), the middle occipital lobe (6.9%) and the precuneus region (5.4%). For the MD values, we obtained a null silver-standard set of 101 voxels. The top regions represented in the null silver-standard are the postcentral gyrus (5.5%) and the lingual gyrus (4.8%).

3 Results

The results are organized as follows. We first show evidence of substantial site effects in the FA and MD maps in Section 3.1, and then show how the different harmonization methods perform at removing those site effects in Section 3.2. In Section 3.3, we discuss the biological variability at each site separately, before and after harmonization and show how site effects affect the number of voxels associations with age. In Section 3.4, we present our experiments for simulating different levels of confounding between age and site. In Section 3.5, we present the replicability of the voxels associated with age for the different harmonization techniques. In Section 3.6, we present the bias in the associations between DTI values and age, and show how the different harmonization techniques perform at correcting for the bias. In Section 3.8, we show how the different harmonization techniques are robust to small sample size studies.

3.1 DTI scalar maps are highly affected by site

In Figure 2a, we present the histogram of FA values for the WM voxels for each participant, stratified by site. We observe a striking systematic difference between the two sites for all values of FA, with an overall difference of 0.07 in the WM (Welch two-sample t-test, $p < 2.2e-16$). Importantly, we notice that the inter-site variability in the histograms is much larger than the intra-site variability, confirming the importance of harmonizing the data across sites. A convenient way to visualize voxel-wise between-site differences in the FA values is plot the average between-site differences as a function of the average across sites. The Bland-Altman plot [Bland and Altman, 1986], also known as the Tukey mean-difference

plot [Cleveland, 1993] or MA-plot [Dudoit et al., 2002, Bolstad et al., 2003] has been used extensively in the genomic literature to compare treatments and investigate dye bias. We use the more common terminology, MA-plot, and present the results for the FA values in Figure 2b. Not surprisingly, the scatterplot is shifted away from the zero line, indicating global site differences. Additionally, there is a large proportion of the voxels (top left voxels) appear to behave differently from other voxels. In the white matter atlas, these voxels are identified as being located in the occipital lobe (middle, inferior and superior gyri, and cuneus), in the fusiform gyrus and in the inferior temporal gyrus. This indicates that the site differences are region-specific, and that a global scaling approach will be most likely insufficient to correct for such local effects.

To further illustrate region-specific differences, we present in Figure 2c the boxplots of FA values for two selected regions, cuneus left and putamen left, stratified by site. **Those results motivate the need of a region-specific harmonization.** We present similar results for the MD maps in Figure 2d,e,f. We note that the site differences appear to be more subtle for MD maps, but nonetheless present. Comparing panel c and panel f, we observe that a brain region exhibiting site differences in FA maps do not necessarily show site differences in the MD maps.

3.2 ComBat successfully removes site effects in DTI scalar maps

In Figure 3, we present the MA-plots before and after each harmonization for the FA maps (see Figure B.4 for the MD maps). While both the scaling and Funnorm methods centered the MA-plots around 0, local site-effects are still apparent. This is consistent with the global nature of the harmonization for these two methods. For the FA maps, RAVEL, SVA and ComBat reduce greatly the inter-site differences. We note that for the MD maps, RAVEL does not seem to account for local site effects. This is not surprising: in Figure B.1, it appears there is a lack of correlation between the average CSF value and average WM value in the MD maps. In other words, the CSF intensities do not act as site surrogates for the WM intensities, and therefore the RAVEL methodology underperforms in this situation. We obtained similar results for the AD and RD maps (Figure B.5 and Figure B.6 respectively).

Next, we calculated a t-statistic at each voxel to measure the association of the DTI scalar values with site. We present in Figure 4a the number of voxels in the WM that are significantly associated with site for each harmonization approach, for both FA and MD maps. A voxel is significant if the p-value calculated from the two-sample t-test is less than 0.05, after correcting for multiple comparisons using Bonferroni correction. Most voxels are associated with site in the absence of harmonization (raw data), and all harmonization methods reduce the number of voxels associated with site for both FA and MD maps at different degree. In agreement with the MA-plots, RAVEL, SVA and ComBat successfully remove site effects for most voxels in the FA maps, but only SVA and ComBat remove site effects for most voxels in the MD maps. Similar results hold for the AD and RD maps (Figure B.8a).

We also calculated t-statistics after summarizing FA and MD values by brain region. Using the Eve template atlas, we identified 156 region of interest (ROIs) overlapping with the WM mask. We present the number of regions significantly associated with site in Figure B.7a.

While all ROIs are associated with site in the absence of harmonization in the FA maps, SVA and ComBat fully remove site effects for all ROIs. Residual site effects are found for more than a third of the ROIs for the Scaling, Funnorm and RAVEL harmonization methods. Similar results hold for the MD maps (140 out of 156 ROIs are affected by site in the absence of normalization).

3.3 Harmonization across sites preserves within-site biological variability

A good harmonization technique should preserve the biological variability at each site separately. To test that, we calculated t-statistics for association with age before harmonization, for site 1 and site 2 separately, and after harmonization, for site 1 and site 2 separately as well. For each harmonization procedure, we computed the Spearman correlation between the unharmonized t-statistics and the harmonized statistics. For Site 1, the correlations are: $\rho = 0.997$ for both Scaling and Funnorm, $\rho = 0.981$ for RAVEL, $\rho = 0.893$ for SVA and $\rho = 0.994$ for ComBat. For Site 2, the correlations are: $\rho = 0.996$ for both Scaling and Funnorm, $\rho = 0.964$ for RAVEL, $\rho = 0.875$ for SVA and $\rho = 0.997$ for ComBat. The ComBat, Scaling and Funnorm methods perform exceptionally well. We note that the correlation is substantially lower for SVA at both sites. This is not surprising; unlike other methods, SVA removes variability that is not associated with age across the whole dataset, but does not protect for the removal of biological variability at each site individually.

In Figure 4b, we present the number of voxels in the WM that are significantly associated with age for each harmonization approach, for both FA and MD maps. Results for the AD and RD maps are presented in Figure B.8b. A voxel is called significant if the p-value calculated from simple linear regression is less than 0.05, after adjusting for multiple comparisons using Bonferroni correction. All harmonization methods increase the number of significant voxels associated with age in comparison to the raw data. ComBat presents the most substantial gain for FA maps (5658 voxels, in comparison to 481 voxels for raw data) and for MD maps (32,203 voxels, in comparison to 23,136 voxels for raw data). Interestingly, we note that for both the AD and RD maps, RAVEL performs as well as ComBat. We also performed a similar analysis at the ROI level: using the white matter atlas, we computed an average FA value at each region, for each participant separately, and subsequently applied the different harmonization techniques; similar results were obtained (see Figure B.7b).

3.4 Harmonization and confounding

In the next sections, we evaluate the performance of the different harmonization procedures by estimating the replicability of the voxels associated with age. We also investigate the robustness of the different harmonization techniques to datasets for which age is confounded by site. The previous results were obtained by harmonizing two sites that were carefully matched for age, gender and ethnicity to minimize potential confounding of those variables with site. However, matching has several limitations. If there is a poor overlap between the covariates of interest across sites, matching will result in a significant exclusion of samples. In addition, the number of scans to be excluded is proportional to the number of covariates to be matched, which can be significant in many applications, making matching infeasible. On the other hand, failing to match for covariates will result in an undesirable situation

where site will be a confounder for the relationship between the DTI values and the phenotype(s) of interest. Thus, a better alternative to matching is to first combine all available data across sites, and then to apply a harmonization technique that is robust to confounding.

Confounding between age and site presents an additional challenge for harmonization, since removing variation associated site can lead to removing variation associated with age if not done carefully. To evaluate the robustness of the different harmonization methods in the presence of statistical confounding between imaging site and age (that is age is unbalanced with respect to site), we selected different subsets of the data to create several confounding scenarios, as shown in Figure 5. For illustration purpose, we chose a voxel in the right thalamus for which the association between FA and age is high. Figure B.3 illustrates the confounding scenarios using median FA values in the WM. We see that for the full data (Figure 5a), the FA values increase linearly with age within each site.

“Positive confounding” and “negative confounding” refer to situations where the relationship between the FA values and age is overestimated and underestimated, respectively, with the same directionality of the true effect. Selecting older samples from Site 2, and younger samples from Site 1, creates positive confounding (Figure 5b). This is because the average FA value for Site 2 is higher than the average value for Site 1. On the other hand, excluding the oldest participants from Site 2 and the youngest participants from Site 1 will create negative confounding (Figure 5c). “Qualitative confounding” is an extreme case of confounding where the estimated direction of the association is reversed with respect to the true association. Selecting younger participants from Site 2, and older participants from Site 1, with no overlap of age between the two sites, creates such confounding (Figure 5d).

We note that in the no-confounding scenario of Figure 5a, the association between the FA values is unbiased in the sense that it is not modified by site. Indeed, the slope using all the data (black line) is similar to the slopes estimated within each site (grey lines). However, the variance of the estimated slope will be inflated due to the unaccounted variation attributable to site.

3.5 ComBat improves the replicability of the voxels associated with age

We evaluated the replicability of the voxels associated with age using the discovery-validation scheme described in Section 2.4. We considered the harmonized dataset as a discovery cohort, and two independent datasets as validation cohorts. We performed a mass-univariate analysis testing for association with age separately for each cohort, and used CAT curves [Irizarry et al., 2005] to measure the replicability of the results between the discovery and validation cohorts. This evaluates the performance of the different harmonization techniques at replicating the voxels associated with age across independent datasets, where replicability is defined as chance that an independent experiment will produce consistent results [Leek and Peng, 2015].

We used two different independent cohorts for estimating the replicability: a larger cohort composed of 292 participants with a similar age distribution (Independent Dataset 1) and a

cohort composed of 105 participants with a slightly older age distribution (Independent Dataset 2), as described in the Methods section. Both cohorts were taken from the PNC [Satterthwaite et al., 2014].

In Figure 6a, we present the CAT curves using Independent Dataset 1 as a validation cohort (same age range). In the absence of confounding (first column), there is good overlap for all methods, including the raw data. ComBat performs the best, with a flat CAT curve around 1. In the positive confounding scenario (second column), all methods perform similar to the raw data, except for the scaling and Funnorm approaches that show substantial inconsistencies with the ranking of the within-site t-statistics, as seen by their CAT curves closer to the diagonal line. This is not surprising; both the scaling and Funnorm approach are global approaches, and because of the positive nature of the confounding, the removal of a global shift associated with site will also remove the global signal associated with age. We note that ComBat performs better than RAVEL and SVA (higher CAT curve).

In the presence of negative confounding and qualitative confounding, combining the data without a proper harmonization technique lead to more severe problems (Figure 6a, third and fourth columns). The CAT curves for the raw data (no harmonization) are considerably below the diagonal line, indicating a negative correlation between the results from the combined dataset and the results from the independent dataset. The negative correlation can be explained by the following: because of the negative (or qualitative) confounding, the t-statistics for the voxels that are truly not associated with age, normally centered around 0, became highly negative because of the site effect. On the other hand, the t-statistics for the voxels associated with age, normally positive for FA, are shifted towards 0. The negative and qualitative confounding render the null voxels significant and create a reversed ranking.

In the negative confounding scenario, all methods, except SVA, are able to recover a ranking that is much more consistent with the true ranking, therefore improving replicability of the results. ComBat yields the highest concordances. In the qualitative confounding situation, only ComBat, Funnorm and the scaling approach improving upon the raw data, with ComBat showing the most substantial improvement. Overall, the results are the most promising for ComBat: the replicability of the top voxels associated with age is dramatically improved for all confounding scenarios, making ComBat a robust harmonization method. Indeed, the ComBat CAT curves are very alike for the four confounding scenarios. The other harmonization approaches show variable performance.

In Figure 6b, we present the CAT curves using another independent dataset, with older participants (Independent Dataset 2). Because we are measuring replicability of the results for two cohorts that have slightly different age ranges, there may be differences in the subset of voxels that are truly associated with age. This can be seen in lower overall concordances curves in Figure 6b. Nevertheless, the results are very consistent with those of Figure 6a. This validation with another additional independent brings more evidence that ComBat performs well at improving the replicability of voxels associated with age, for all confounding scenarios.

In Appendix A.2, we compare the performance of ComBat on confounded subsamples vs matched subsamples at replicating associations between age and FA. The results show that ComBat provides a useful level of correction even in the presence of confounding, comparable to balanced samples.

3.6 ComBat successfully recovers the true effect sizes

In this section, we evaluate the bias in the estimated changes in FA associated with age ($\hat{\beta}_{\text{age}}\text{FA}$) for each harmonization procedure, for the different confounding scenarios. We refer to $\hat{\beta}_{\text{age}}\text{FA}(v)$ as the estimated “effect size” for voxel v . The effect size can be estimated using linear regression (slope coefficient associated with age). In principle, to assess unbiasedness, we would need to know the true effect sizes $\beta_{\text{age}}\text{FA}$. We circumvent this by estimating the effect sizes on the signal silver-standard described in Section 2.5. For each site, we calculated the effect size for each voxel of the signal silver-standard by running a simple linear regression for age, and retaining the regression coefficient for age as the estimated effect size. We took the average across the two sites at each voxel as the estimated true effect size. This resulted in a distribution of 2265 effect sizes for the signal voxels, with a median effect size close to 0.004, presented in the left boxplot of Figure B.10a. We also estimated the true effect sizes for voxels not associated with age (null voxels described in Section 2.5). We obtained a distribution of 1932 effect sizes for the null signal. Not surprisingly, those effects sizes are roughly centered at 0 (right boxplot, Figure B.10a).

In Figure 7a, we present the distribution of the estimated effect sizes on the signal silver-standard for all methods, and for all confounding scenarios. The dashed lined represents the median effect size of the true effect sizes, and the solid line represents an effect size of 0. As expected, the effect sizes in the raw data (datasets combined without harmonization) are consistent with the type of confounding; positive confounding shifts the effect sizes positively, and the negative and qualitative confounding shifts the effect sizes negatively. ComBat is the only harmonization technique that fully recovers the true effect sizes for all confounding scenarios in terms of median value and variability. Funnorm and RAVEL both reduced the bias in the effect sizes distribution, and both underestimate the true associations. We note that RAVEL performs sensibly worse for the qualitative confounding scenario. Interestingly, SVA does not achieve any bias correction for any of the confounding scenarios; the distribution of the estimated effect sizes resemble those of the unharmonized dataset. This could be explained by the fact that SVA method “protects” for the present association between the outcome and the covariate of interest, and therefore an association that is biased in the original dataset will remain biased in the SVA-corrected dataset. Similarly, we present in Figure 7b the distribution of the estimated effect sizes for the null silver-standard. We recall that a successful harmonization approach will result in a boxplot centered around 0. The results are similar to Figure 7a; ComBat successfully recovers the true effect size distribution for all confounding scenarios. Results for MD maps are presented in Figure B.10b and Figure B.11.

The retrieval of unbiased effect sizes for both the signal and the null silver-standard strongly suggests that ComBat successfully removed the site effect in the combined datasets without removing the signal associated with age, even in the presence of substantial confounding

between age and site. The FA changes estimated after ComBat for voxels highly associated with age are similar to the FA changes measured at each site separately.

3.7 ComBat improves statistical power

In Figure 8, we present the distribution of the WM voxels-wise t-statistics measuring association with age in the FA maps for four combinations of the data: Site 1 and Site 2 analyzed separately, Site 1 and Site 2 combined without harmonization, and Site 1 and Site 2 combined and harmonized with ComBat. The goal of combining datasets from different sites is to increase the sample size, and therefore the power of the statistical analysis. We therefore expect t-statistics with higher magnitude for voxels truly associated with age. Moreover, we note that most of the t-statistics will be positive as a consequence of the global increase in FA associated with development of the brain in teenagers [Tamnes et al., 2010, Bava et al., 2010, Lebel and Beaulieu, 2011].

In Figure 8a, in which we present the t-statistics for all voxels in the WM, we observe an opposite effect. The distribution of the t-statistics for the two sites combined without harmonization is shifted towards 0 (mean t-statistic of 1.4) in comparison to the t-statistics obtained from both sites separately (mean t-statistic of 1.7 and 2.3 for site 1 and site 2 respectively). This strongly indicates that combining data from multiple sites, without harmonization, is counter-productive and impairs the quality of the data. On the other hand, combining and harmonizing data with ComBat results in a distribution of higher t-statistics on average (mean t-statistic of 2.8). We present in Figure 8b the t-statistics in template space with and without ComBat.

To further examine the effects of harmonization on the data, we present the distribution of the t-statistics for voxels that are truly associated with age (signal silver-standard described in Section 2.5) in Figure 8c, and voxels that are truly not associated with age (null silver-standard described in Section 2.5) in Figure 8d). This confirms that ComBat increases the statistical power at finding voxels truly associated with age, as seen by the distribution of t-statistics substantially shifted to the right in Figure 8c. The mean t-statistic for the raw data and after ComBat is 4.3 and 8.3 respectively. ComBat also keeps the t-statistics of the null voxels tightly centered around 0 (Figure 8d). In Figure 8e, we present the number of voxels significantly associated with age ($p < 0.05$) after adjusting for multiple comparisons using Bonferroni correction. The results strengthen our observations that harmonization is needed in order to successfully combine multi-site data.

We present the results for the MD maps in Figure B.9. It is expected to observe many voxels showing a negative association between MD and age in teenagers [Tamnes et al., 2010, Bava et al., 2010, Lebel and Beaulieu, 2011], and therefore to observe a distribution of t-statistics shifted towards negative values (as opposed to the t-statistics distribution in FA maps). Again, ComBat successfully increases the magnitude of the t-statistics for the signal voxels (distribution of the t-statistics highly shifted away from 0 in Figure B.9c), while maintaining the t-statistics for the null voxels centered around 0 (Figure B.9d).

3.8 ComBat is robust to small sample sizes

A major advantage of ComBat over other methods is the use of Empirical Bayes to improve the estimation and removal of the site effects in small sample size settings. To assess the robustness of the different harmonization approaches for combining small samples size studies, we created $B = 100$ random subsets of size $n = 20$ across sites. Specifically, we selected for each subset 10 participants at random from each site. For each subset, we applied the different harmonization methods and calculated voxel-wise t-statistics in the WM, for testing the association of the FA values with age, for a total of 100 t-statistic maps. To obtain an estimated gold-standard for a t-statistic map obtained with studies of sample size 20, that we refer to as a silver-standard, we created $B = 100$ random subsets of size 20 from site 1, and $B = 100$ additional random subsets of size 20 from site 2. Because subsets are created within site, they are not affected by site effects and results obtained from those subsets should be superior or as good as any of the results obtained from the harmonized subsets.

In Figure 9a, we present the average CAT curve for each harmonization method (average taken across the random subsets) together with the silver-standard CAT curve (dark blue), for the FA maps. All methods improve the replicability of the voxels associated with age. We note that ComBat performs as well as the silver-standard, successfully removing most of the site effects. In Figure 9b, we present the densities of the t-statistics for the top voxels associated with age (signal voxels described in Section 2.5) for the FA maps. We note that all methods improve the magnitude of the t-statistics, therefore increasing statistical power, with ComBat showing the best performance, notably performing as well as the silver-standard. In Figure 9c, we present the densities of the t-statistics for voxels not associated with age (null voxels described in Section 2.5) for the FA maps; a good harmonization method should result in t-statistics centered around 0. The global scaling approach, functional normalization and ComBat correctly return t-statistics centered around 0 that are similar to the silver-standard. SVA and RAVEL do not perform as well (densities shifted away from 0). Overall, the results show that ComBat is a very promising harmonization method even for small sample size studies, doing as well as a dataset that was not affected by site effects. Similar results were obtained for the MD maps, presented in the panels (d–f) of Figure 9.

In Appendix A.3, we investigate the stability of the ComBat harmonization parameters by running ComBat on random subsamples of size $m \in \{10, 20, \dots, 210\}$. repeat the subsampling $B = 100$ times. We obtained that site effects estimated from subsamples approximate well the site effects estimated from the full dataset.

4 Discussion

In this work, we investigated the effects of combining DTI studies across sites and scanners on the statistical analyses. We used FA and MD maps from data acquired at two sites with different scanners. We first showed that combining the two studies without proper harmonization led to a decrease in power of detecting voxels associated with age. This confirmed that DTI measurements are highly affected by small changes in the scanner parameters, as those affect the underlying water diffusivity. This motivated the need for

harmonizing data across sites and scanners. We then adapted and compared several statistical harmonization techniques for DTI studies.

Using a comprehensive evaluation framework that respects the importance of biological variation in the data, we showed that ComBat, a popular batch effect correction tool used in genomics, performs the best at harmonizing FA and MD maps. It allows site effects to be location-specific, but pools information across voxels to improve the statistical estimation of the site effects. More specifically, we showed that ComBat substantially increases the replicability of the voxels associated with age across independent experiments. We also investigated the robustness of the proposed harmonization methods when the associations of age and DTI measurements are confounded by site as a consequence of possible unbalanced data, as well as robustness to small sample sizes. ComBat was the best at improving the results across all scenarios, and appeared to be robust to small sample size studies. Indeed, it was able to recover the true associations between the FA (and MD) values and age, despite the bias introduced by the association between site and age.

Global scaling and functional normalization [Fortin et al., 2014] did not perform well overall. This is not surprising; both of these histogram-normalization methods fail to account for the spatial heterogeneity of the site effects throughout the brain. We also compared ComBat to RAVEL, an intensity normalization technique previously proposed for T1-w images [Fortin et al., 2016a]. RAVEL performed well for the FA maps, for which the FA values in the CSF reflect the technical variation in the WM. However, RAVEL did not perform well for the MD maps; the site effects in the CSF were not correlated with the site effects in the WM. We also compared ComBat to SVA [Leek and Storey, 2007, 2008], an algorithm developed for genomics data that estimates unwanted variation that is orthogonal to the biological variation. SVA was successful at estimating and removing the site effects, but did not perform as well as ComBat for datasets for which age was confounded with site.

The ComBat methodology can be extended in several ways. In the case of a categorical outcome, for instance disease status, one could estimate the site effect parameters using only participants from a healthy population. This approach would be particularly useful for combining clinical studies with heterogeneous disease effects, such as ASD and traumatic brain injury (TBI). Indeed, for small sample sizes, distinguishing between disease heterogeneous effects and site effects might be intractable, and using a relatively more stable healthy population for normalizing the data has been shown to improve performance Linn et al. [2016b].

Future ComBat models might also draw strength from spatial correlation by spatially restraining the estimation of the hyperparameters for the prior distributions to only pool information across neighboring voxels. Another extension would be to incorporate an inverse probability weighting (IPW) scheme to explicitly model statistical confounding between the phenotype of interest in site. IPW has been shown to improve results when there is presence of confounding in imaging studies [Linn et al., 2016a], especially in mitigating multivariate confounding for prediction.

It is also possible to extend ComBat for longitudinal studies. For such studies, it is common for scanners to undergo software upgrades. In addition, scans from follow-up visits can be acquired on a different scanner. Those changes are likely to add unwanted variability to the brain trajectories, and in certain cases to cancel out subtle phenotypic effects associated with time. If time is not entirely confounded with scanner, it is possible to remove these undesirable scanner effects by adding a time variable to the ComBat model as an additional covariate to adjust for. This will make sure that true longitudinal changes in the brain are preserved while scanner effects are removed.

While this paper has focused on the harmonization of imaging data across sites and scanners, another important challenge is the harmonization of imaging data within a site. Indeed, even for scans acquired on the same scanner, between-participant unwanted variation that is technical in nature also exists. This requires a harmonization technique that is not dependent on a site, or scanner, variable. In genomics, latent factor approaches that estimate unknown source of variation have been successfully used, such as SVA [Leek and Storey, 2007] and RUV [Gagnon-Bartsch and Speed, 2012]. Recently, a similar approach called RAVEL [Fortin et al., 2016a] has been developed for the harmonization of structural MRI intensities using a control region for the estimation of the latent factors of unwanted variation [Fortin et al., 2016a]. Similar to RAVEL, the ComBat framework can be easily extended to within-site harmonization by estimating latent factors of unwanted variation from a control region. Indeed, the latent factors estimated from a control region can be integrated as within-site location parameters in the ComBat model presented in Equation 2. The choice of an appropriate control region for DTI studies is part of our future work.

We also note that ComBat methodology is readily applicable to tract-based spatial statistics (TBSS) [Smith et al., 2006] analyses; such analyses are part of our future work. Although we have shown the performance of ComBat in the context of DTI scalar maps, the ComBat model is widely applicable beyond this setting. The ComBat model does not make any assumptions regarding the neuroimaging technique being used, therefore making it applicable to other imaging techniques. For instance, it can be used to harmonize connectivity data across different processing protocols, such as seed-based connectivity maps in resting-state fMRI or measures of structural connectivity derived from DTI. In addition, while we used voxels as features to be harmonized in the ComBat model, the ComBat algorithm can be applied to measurements summarized at the ROI level, making ComBat a promising harmonization technique for volumetric and cortical thickness studies.

5 Software

All of the postprocessing analysis was performed in the R statistical software (v3.2.0). For SVA and ComBat, reference implementations from the *sva* package were used (v3.22.0). All figures were generated in R with customized and reproducible scripts, using several functions from the package *fslr* [Muschelli et al., 2015] (v2.12). We have adapted and implemented the ComBat methodology to imaging data, and the software is available in both R and Matlab on GitHub (<https://github.com/Jfortin1/ComBatHarmonization>).

Acknowledgments

Funding

The research was supported in part by R01NS085211 and R21NS093349 from the National Institute of Neurological Disorders and Stroke, R01MH092862 and R01MH107703 from the National Institute of Mental Health and R01HD089390 from the National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Abbreviations

ADNI	Alzheimer's Disease NeuroImaging Initiative
AD	Axial diffusivity
CAT	Concordance at the top
ComBat	Combatting batch effects when combining batches of gene expression microarray data
CoV	Coefficient of variation
CSF	Cerebrospinal fluid
DTI	Diffusion tensor imaging
EB	Empirical Bayes
FA	Fractional anisotropy
GM	Grey matter
GS	Global scaling
IBMA	Image-based meta analysis
IPW	Inverse probability weighting
MD	Mean diffusivity
MRI	Magnetic resonance imaging
OLS	Ordinary least squares
RD	Radial diffusivity
RAVEL	Removal of artificial voxel effect by linear regression
RMSE	Root mean square error
RISH	Rotation invariant spherical harmonic
ROI	Region of interest
SVA	Surrogate variable analysis
SVD	Singular value decomposition

T1-w	T1-weighted
TBI	Traumatic brain injury
TBSS	Tract-based spatial statistics
TDC	Typically developing control
WM	White matter
WMPM	White matter parcellation map

References

- Alexander, Andrew L., Hasan, Khader M., Lazar, Mariana, Tsuruda, Jay S., Parker, Dennis L. Analysis of partial volume effects in diffusion-tensor mri. *Magnetic Resonance in Medicine*. 2001; 45(5): 770–780. [PubMed: 11323803]
- Alexander, Andrew L., Lee, Jee Eun, Lazar, Mariana, Field, Aaron S. Diffusion tensor imaging of the brain. *Neurotherapeutics*. 2007; 4(3):316–329. [PubMed: 17599699]
- Ashtari, Manzar, Cervellione, Kelly L., Hasan, Khader M., Wu, Jinghui, McIlree, Carolyn, Kester, Hana, Ardekani, Babak A., Roofeh, David, Szeszko, Philip R., Kumra, Sanjiv. White matter development during late adolescence in healthy males: a cross-sectional diffusion tensor imaging study. *Neuroimage*. 2007; 35(2):501–510. [PubMed: 17258911]
- Barnea-Goraly, Naama, Menon, Vinod, Eckert, Mark, Tamm, Leanne, Bammer, Roland, Karchemskiy, Asya, Dant, Christopher C., Reiss, Allan L. White matter development during childhood and adolescence: a cross-sectional diffusion tensor imaging study. *Cerebral cortex*. 2005; 15(12):1848–1854. [PubMed: 15758200]
- Bava, Sunita, Thayer, Rachel, Jacobus, Joanna, Ward, Megan, Jernigan, Terry L., Tapert, Susan F. Longitudinal characterization of white matter maturation during adolescence. *Brain research*. 2010; 1327:38–46. [PubMed: 20206151]
- Bland Martin J, Altman Douglas G. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*. 1986; 327(8476):307–310.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003; 19(2):185–193. [PubMed: 12538238]
- Cleveland, W. Visualizing data. at & t bell laboratories; murray hill nj: 1993.
- Cleveland, William S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*. 1979; 74:829–836.
- Cleveland, William S. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*. 1981; 35(1):54.
- Correia, Marta Morgado, Carpenter, Thomas A., Williams, Guy B. Looking for the optimal dti acquisition scheme given a maximum scan time: are more b-values a waste of time? *Magnetic resonance imaging*. 2009; 27(2):163–175. [PubMed: 18687552]
- De Wit, Stella J., Alonso, Pino, Schweren, Lizanne, Mataix-Cols, David, Lochner, Christine, Menchón, José M., Stein, Dan J., Fouche, Jean-Paul, Soriano-Mas, Carles, Sato, Joao R., et al. Multicenter voxel-based morphometry mega-analysis of structural brain scans in obsessive-compulsive disorder. *American journal of psychiatry*. 2014; 171(3):340–349. [PubMed: 24220667]
- Dudoit, Sandrine, Yang, Yee Hwa, Callow, Matthew J., Speed, Terence P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica*. 2002:111–139.
- Fortin, Jean-Philippe, Labbe, Aurelie, Lemire, Mathieu, Zanke, Brent, Hudson, Thomas, Fertig, Elana, Greenwood, Celia, Hansen, Kasper D. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*. 2014; 15(11):503.doi: 10.1186/s13059-014-0503-2 [PubMed: 25599564]

- Fortin, Jean-Philippe, Sweeney, Elizabeth M., Muschelli, John, Crainiceanu, Ciprian M., Shinohara, Russell T. Alzheimer's Disease Neuroimaging Initiative, et al. Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*. 2016a; 132:198–212. [PubMed: 26923370]
- Fortin, Jean-Philippe, Triche, Timothy J., Hansen, Kasper D. Preprocessing, normalization and integration of the illumina humanmethylationepic array with minfi. *Bioinformatics*. 2016b:btw691.
- Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 2012; 13(3):539–552. DOI: 10.1093/biostatistics/kxr034 [PubMed: 22101192]
- Garyfallidis, Eleftherios, Brett, Matthew, Amirbekian, Bagrat, Rokem, Ariel, Van Der Walt, Stefan, Descoteaux, Maxime, Nimmo-Smith, Ian. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*. 2014; 8:8. [PubMed: 24600385]
- Ghanbari, Yasser, Smith, Alex R., Schultz, Robert T., Verma, Ragini. Identifying group discriminative and age regressive sub-networks from dti-based connectivity via a unified framework of non-negative matrix factorization and graph embedding. *Medical image analysis*. 2014; 18(8):1337–1348. [PubMed: 25037933]
- Giannelli, Marco, Cosottini, Mirco, Michelassi, Maria Chiara, Lazzarotti, Guido, Belmonte, Gina, Bartolozzi, Carlo, Lazzeri, Mauro. Dependence of brain dti maps of fractional anisotropy and mean diffusivity on the number of diffusion weighting directions. *Journal of Applied Clinical Medical Physics*. 2009; 11(1)
- Giorgio, Antonio, Watkins, KE., Chadwick, Martin, James, S., Winmill, Louise, Douaud, Gwenaëlle, De Stefano, Nicola, Matthews, Paul M., Smith, Steve M., Johansen-Berg, Heidi, et al. Longitudinal changes in grey and white matter during adolescence. *Neuroimage*. 2010; 49(1):94–103. [PubMed: 19679191]
- Huo, Yuankai, Aboud, Katherine, Kang, Hakmook, Cutting, Laurie E., Landman, Bennett A. Mapping lifetime brain volumetry with covariate-adjusted restricted cubic spline regression from cross-sectional multi-site mri. *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer; 2016. p. 81-88.
- Irizarry, Rafael A., Warren, Daniel, Spencer, Forrest, Kim, Irene F., Biswal, Shyam, Frank, Bryan C., Gabrielson, Edward, Garcia, Joe GN., Geoghegan, Joel, Germino, Gregory, Griffin, Constance, Hilmer, Sara C., Hoffman, Eric, Jedlicka, Anne E., Kawasaki, Ernest, Martínez-Murillo, Francisco, Morsberger, Laura, Lee, Hannah, Petersen, David, Quackenbush, John, Scott, Alan, Wilson, Michael, Yang, Yanqin, Ye, Shui Qing, Yu, Wayne. Multiple-laboratory comparison of microarray platforms. *Nature Methods*. 2005; 2(5):345–50. DOI: 10.1038/nmeth756 [PubMed: 15846361]
- Jahanshad, Neda, Kochunov, Peter V., Sprooten, Emma, Mandl, René C., Nichols, Thomas E., Almasy, Laura, Blangero, John, Brouwer, Rachel M., Curran, Joanne E., de Zubicaray, Greig I., et al. Multisite genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the enigma-dti working group. *Neuroimage*. 2013; 81:455–469. [PubMed: 23629049]
- Jenkinson, Mark, Smith, Stephen. A global optimisation method for robust affine registration of brain images. *Medical image analysis*. 2001; 5(2):143–156. [PubMed: 11516708]
- Jenkinson, Mark, Bannister, Peter, Brady, Michael, Smith, Stephen. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*. 2002; 17(2):825–841. [PubMed: 12377157]
- Evan Johnson W, Li Cheng, Rabinovic Ariel. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*. 2007; 8(1):118–127. [PubMed: 16632515]
- Jones, Derek K. The effect of gradient sampling schemes on measures derived from diffusion tensor mri: a monte carlo study. *Magnetic Resonance in Medicine*. 2004; 51(4):807–815. [PubMed: 15065255]
- Kim, Mina, Ronen, Itamar, Ugurbil, Kamil, Kim, Dae-Shik. Spatial resolution dependence of dti tractography in human occipito-callosal region. *Neuroimage*. 2006; 32(3):1243–1249. [PubMed: 16861009]
- Klein, Kathleen Oros, Grinek, Stepan, Bernatsky, Sasha, Bouchard, Luigi, Ciampi, Antonio, Colmegna, Ines, Fortin, Jean-Philippe, Gao, Long, Hivert, Marie-France, Hudson, Marie, et al. funtoonorm: an r package for normalization of dna methylation data when there are multiple cell or tissue types. *Bioinformatics*. 2015:btv615.

- Kochunov, Peter, Jahanshad, Neda, Sprooten, Emma, Nichols, Thomas E., Mandl, René C., Almasy, Laura, Booth, Tom, Brouwer, Rachel M., Curran, Joanne E., de Zubicaray, Greig I., et al. Multisite study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. *NeuroImage*. 2014; 95:136–150. [PubMed: 24657781]
- Krogsrud, Stine K., Fjell, Anders M., Tamnes, Christian K., Grydeland, Håkon, Mork, Lia, Due-Tønnessen, Paulina, Bjørnerud, Atle, Sampaio-Baptista, Cassandra, Andersson, Jesper, Johansen-Berg, Heidi, et al. Changes in white matter microstructure in the developing brain: a longitudinal diffusion tensor imaging study of children from 4 to 11 years of age. *NeuroImage*. 2016; 124:473–486. [PubMed: 26375208]
- Lebel, Catherine, Beaulieu, Christian. Longitudinal development of human brain wiring continues from childhood into adulthood. *The Journal of Neuroscience*. 2011; 31(30):10937–10947. [PubMed: 21795544]
- Lebel, Catherine, Gee, M., Camicioli, R., Wieler, M., Martin, W., Beaulieu, Christian. Diffusion tensor imaging of white matter tract evolution over the lifespan. *Neuroimage*. 2012; 60(1):340–352. [PubMed: 22178809]
- Leek, Jeffrey T., Peng, Roger D. Opinion: Reproducible research can still be wrong: adopting a prevention approach. *Proc Natl Acad Sci U S A*. Feb; 2015 112(6):1645–6. DOI: 10.1073/pnas.1421412111 [PubMed: 25670866]
- Leek, Jeffrey T., Storey, John D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*. 2007; 3(9):1724–1735. DOI: 10.1371/journal.pgen.0030161 [PubMed: 17907809]
- Leek, Jeffrey T., Storey, John D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*. 2008; 105(48):18718–18723. DOI: 10.1073/pnas.0808709105
- Leek, Jeffrey T., Evan Johnson, W., Parker, Hilary S., Jaffe, Andrew E., Storey, John D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012; 28(6):882–883. DOI: 10.1093/bioinformatics/bts034 [PubMed: 22257669]
- Linn, Kristin A., Gaonkar, Bilwaj, Doshi, Jimit, Davatzikos, Christos, Shinohara, Russell T. Addressing confounding in predictive models with an application to neuroimaging. *The international journal of biostatistics*. 2016a; 12(1):31–44. [PubMed: 26641972]
- Linn, Kristin A., Gaonkar, Bilwaj, Satterthwaite, Theodore D., Doshi, Jimit, Davatzikos, Christos, Shinohara, Russell T. Control-group feature normalization for multivariate pattern analysis of structural mri data using the support vector machine. *NeuroImage*. 2016b; 132:157–166. [PubMed: 26915498]
- Mirzaalian H, Ning L, Savadjiev P, Pasternak O, Bouix S, Michailovich O, Grant G, Marx CE, Morey RA, Flashman LA, et al. Inter-site and inter-scanner diffusion mri data harmonization. *NeuroImage*. 2016; 135:311–323. [PubMed: 27138209]
- Muschelli, John, Sweeney, Elizabeth M., Lindquist, Martin A., Crainiceanu, Ciprian M. fslr: Connecting the fsl software with r. *The R Journal*. Feb; 2015 7(1):163–175. [PubMed: 27330830]
- Ou, Yangming, Sotiras, Aristeidis, Paragios, Nikos, Davatzikos, Christos. Dramms: Deformable registration via attribute matching and mutual-saliency weighting. *Medical image analysis*. 2011; 15(4):622–639. [PubMed: 20688559]
- Parmigiani, Giovanni, Garrett, Elizabeth S., Irizarry, Rafael A., Zeger, Scott L. The analysis of gene expression data. Springer; 2003. The analysis of gene expression data: an overview of methods and software; p. 1-45.
- Salimi-Khorshidi, Gholamreza, Smith, Stephen M., Keltner, John R., Wager, Tor D., Nichols, Thomas E. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage*. 2009; 45(3):810–823. [PubMed: 19166944]
- Satterthwaite, Theodore D., Elliott, Mark A., Ruparel, Kosha, Loughhead, James, Prabhakaran, Karthik, Calkins, Monica E., Hopson, Ryan, Jackson, Chad, Keefe, Jack, Riley, Marisa, et al. Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage*. 2014; 86:544–553. [PubMed: 23921101]

- Schmithorst, Vincent J., Wilke, Marko, Dardzinski, Bernard J., Holland, Scott K. Correlation of white matter diffusivity and anisotropy with age during childhood and adolescence: A cross-sectional diffusion-tensor mr imaging study 1. *Radiology*. 2002; 222(1):212–218. [PubMed: 11756728]
- Shinohara, Russell T., Sweeney, Elizabeth M., Goldsmith, Jeff, Shiee, Navid, Mateen, Farrah J., Calabresi, Peter A., Jarso, Samson, Pham, Dzung L., Reich, Daniel S., Crainiceanu, Ciprian M. Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, and Alzheimer's Disease Neuroimaging Initiative. Statistical normalization techniques for magnetic resonance imaging. *Neuroimage Clin*. 2014; 6:9–19. DOI: 10.1016/j.nicl.2014.08.008 [PubMed: 25379412]
- Smith, Stephen M. Fast robust automated brain extraction. *Hum Brain Mapp*. Nov; 2002 17(3):143–55. DOI: 10.1002/hbm.10062 [PubMed: 12391568]
- Smith, Stephen M., Jenkinson, Mark, Johansen-Berg, Heidi, Rueckert, Daniel, Nichols, Thomas E., Mackay, Clare E., Watkins, Kate E., Ciccarelli, Olga, Zaheer Cader, M., Matthews, Paul M., et al. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*. 2006; 31(4):1487–1505. [PubMed: 16624579]
- Tamnes, Christian K., Østby, Ylva, Fjell, Anders M., Westlye, Lars T., Due-Tønnessen, Paulina, Walhovd, Kristine B. Brain maturation in adolescence and young adulthood: regional age-related changes in cortical thickness and white matter volume and microstructure. *Cerebral cortex*. 2010; 20(3):534–548. [PubMed: 19520764]
- Tristán-Vega, Antonio, Aja-Fernández, Santiago. Dwi filtering using joint information for dti and hardi. *Medical image analysis*. 2010; 14(2):205–218. [PubMed: 20005152]
- Turner, Jessica A. The rise of large-scale imaging studies in psychiatry. *Giga Science*. 2014; 3(1):29. [PubMed: 25793106]
- Westlye, Lars T., Walhovd, Kristine B., Dale, Anders M., Bjørnerud, Atle, Due-Tønnessen, Paulina, Engvig, Andreas, Grydeland, Håkon, Tamnes, Christian K., Østby, Ylva, Fjell, Anders M. Lifespan changes of the human brain white matter: diffusion tensor imaging (dti) and volumetry. *Cerebral cortex*. 2009:bhp280.
- Zhan, Liang, Leow, Alex D., Jahanshad, Neda, Chiang, Ming-Chang, Barysheva, Marina, Lee, Agatha D., Toga, Arthur W., McMahon, Katie L., de Zubicaray, Greig I., Wright, Margaret J., et al. How does angular resolution affect diffusion imaging measures? *Neuroimage*. 2010; 49(2):1357–1371. [PubMed: 19819339]
- Zhang Y, Brady M, Smith S. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*. Jan; 2001 20(1): 45–57. DOI: 10.1109/42.906424 [PubMed: 11293691]
- Zhu, Tong, Liu, Xiaoxu, Gaugh, Michelle D., Connelly, Patrick R., Ni, Hongyan, Ekholm, Sven, Schifitto, Giovanni, Zhong, Jianhui. Evaluation of measurement uncertainties in human diffusion tensor imaging (dti)-derived parameters and optimization of clinical dti protocols with a wild bootstrap analysis. *Journal of Magnetic Resonance Imaging*. 2009; 29(2):422–435. [PubMed: 19161198]
- Zhu, Tong, Hu, Rui, Qiu, Xing, Taylor, Michael, Tso, Yuen, Yiannoutsos, Constantin, Navia, Bradford, Mori, Susumu, Ekholm, Sven, Schifitto, Giovanni, et al. Quantification of accuracy and precision of multi-center dti measurements: a diffusion phantom and human brain study. *Neuroimage*. 2011; 56(3):1398–1411. [PubMed: 21316471]

Appendix A

A.1 Generalization of ComBat for nonlinear covariate effects

In Section 2.3.5, we described the location/scale model used in ComBat where each individual covariate is expected to be linearly associated with the DTI scalar measurements. Here, we extend the model for nonlinearities. This is motivated by the problem of combining multi-site data for the study of life-span trajectories, for which it is common to observe nonlinear associations between age and imaging measurements [Westlye et al., 2009, Lebel et al., 2012].

Let the data come from m imaging sites, containing each n_i scans within site i for $i = 1, 2, \dots, m$, for voxel $v = 1, 2, \dots, p$. Let y_{ijv} represent the FA measure for voxel v for scan j for site i . Let $n = \sum_{i=1}^m n_i$ be the total number of scans. A more general location/scale model is the following:

$$y_{ijv} = \alpha_v + f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v) + \gamma_{iv} + \delta_{iv} \varepsilon_{ijv}, \quad (\text{A.1})$$

where α_v is the overall FA measure for voxel v , \mathbf{X} is the $n \times K$ design matrix for the K covariates of interest (e.g. gender, age), f is a prespecified multivariate function of the covariates parametrized by $\boldsymbol{\beta}_v$. We assume that the form of f is the same for all voxels, and that $\boldsymbol{\beta}_v$ is sufficient to capture voxel-specific effects. The terms γ_{iv} and δ_{iv} represent the additive and multiplicative site effects of site i for voxel v , respectively. We assume that the error terms ε_{ijv} have mean zero and variance σ_v^2 . We note that in the original ComBat model, $f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v) = \mathbf{X}_{ij} \boldsymbol{\beta}_v$ was chosen and the parameter-vector $\boldsymbol{\beta}_v$ was estimated using ordinary least squares (OLS).

Once the component $f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v)$ is estimated from the data, the ComBat methodology can be applied as before on the residuals, and the harmonized values are defined as

$$y_{ijv}^{\text{ComBat}} = \frac{y_{ijv} - \hat{\alpha}_v - f(\mathbf{X}_{ij}, \hat{\boldsymbol{\beta}}_v) - \gamma_{iv}^*}{\delta_{iv}^*} + \hat{\alpha}_v + f(\mathbf{X}_{ij}, \hat{\boldsymbol{\beta}}_v).$$

A popular choice for the modeling of the nonlinear component $f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v)$ is to use cubic splines. Cubic splines are composed of piecewise third-order polynomials with control points (knots) specified in advance. They allow to model nonlinear relationships between two variables in a flexible and smooth fashion. For instance, they have been previously used to model nonlinear relationship between age and brain volumetric measurements [Huo et al., 2016]. Let $\{N_l(x_k)\}_{l=1}^{L_k}$ denote the collection of functions for a given cubic spline basis for the k -th covariate x_k . For each voxel v , we can write $f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v)$ as

$$f(\mathbf{X}_{ij}, \boldsymbol{\beta}_v) = \sum_{k=1}^K \mathbf{N}(x_k) \boldsymbol{\beta}_{vk}$$

where $\mathbf{N}(x_k)$ denotes the $L_k \times n$ design matrix for the basis functions for the k -th covariate, and $\boldsymbol{\beta}_{vk}$ the $L_k \times 1$ vector of coefficients for $\mathbf{N}(x_k)$. We can then rewrite Equation A.1 as

$$y_{ijv} = \alpha_v + \sum_{k=1}^K \mathbf{N}(x_k) \boldsymbol{\beta}_{vk} + \gamma_{iv} + \delta_{iv} \varepsilon_{ijv} \quad (\text{A.2})$$

and estimate $\sum_{k=1}^K N(x_k) \beta_{vk}$ using usual OLS.

A.2 Analysis of matched versus confounded subsamples

In this appendix, we compare the CAT plots from the raw and ComBat-harmonized data in confounded subsamples to CAT plots obtained from matched subsamples. For each confounding scenario presented in the Results section (positive, negative and qualitative confounding), we created a random subsample for which age is matched across the two sites. We made sure that the size of this matched subsample is similar to the size of the confounded subsample (n=117 for positive confounding, n=187 for negative confounding and n=115 for qualitative confounding). For each scenario, we obtained CAT curves using the two independent datasets as validation cohorts. We present in Figure A.1 the CAT curves for the matched subsamples (solid lines) and for the confounded subsamples (dotted lines). A successful ComBat harmonization will result in the two ComBat curves (light blue) close to each other. We observe good performance of ComBat for all confounding scenarios. For the negative and qualitative cases, there is a remarkable improvement upon the raw data (dotted blue curves compared to dotted black curves), and the ComBat-corrected curves for those confounded subsamples are almost as good as the ComBat-corrected curves for the matched subsamples. For positive confounding, ComBat does improve upon the raw data, but the CAT curve is a bit lower in the independent dataset 1 (dotted blue line compared to solid blue line). Overall, this shows that ComBat results in a useful level of correction in presence of confounding.

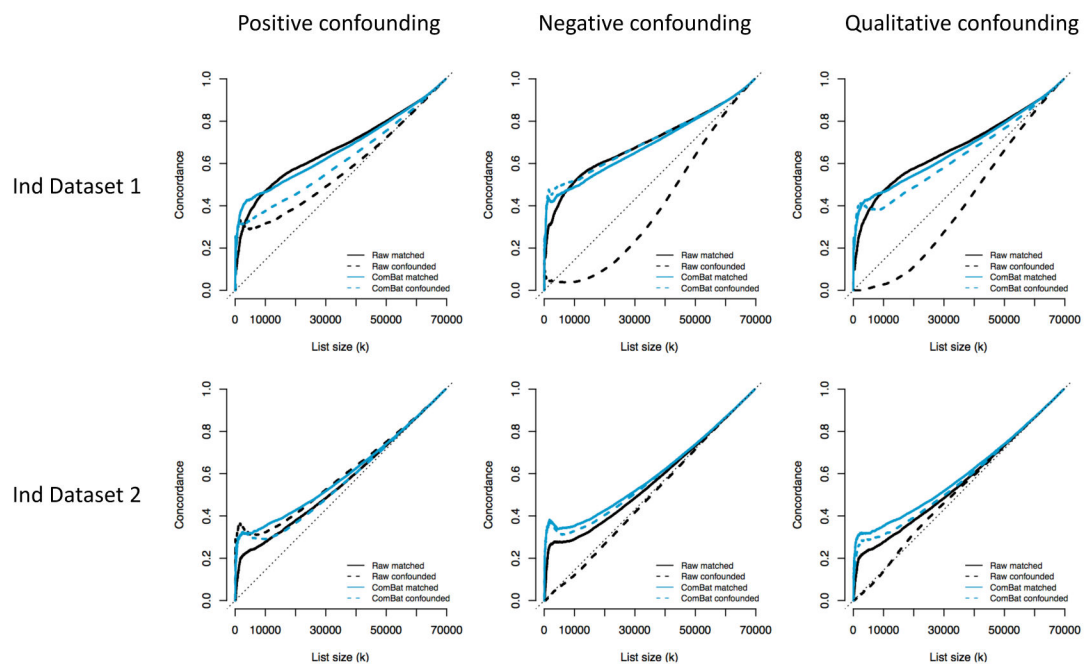


Figure A.1. CAT plots for confounded and matched subsamples

For each confounding scenario, the solid lines represent represent the CAT curves for subsamples of the full dataset that are matched for age across the two sites (no

confounding). The sample size of the subsamples are similar to the sample sizes of the confounded subsamples. The CAT curves for the confounded subsamples are represented by the dotted lines, and correspond to the CAT curves described in the Results section of the manuscript. For the first row, the validation dataset was the Independent dataset 1 (similar age range). For the second row, the validation dataset was the Independent dataset 2, with a slightly older age range.

A.3 Stability analysis of the ComBat harmonization

To investigate the stability of the ComBat harmonization parameters, we ran the ComBat algorithm $B = 100$ times by randomly selecting subsamples of size m from the full dataset, on the FA values. We considered the values $m \in \{10, 20, \dots, 210\}$. We made sure to select the same number of participants from each of the two sites to create balanced subsamples. For each value of m , we estimated the site effects for both sites using ComBat, and averaged the site effects across voxels. We present in Figure A.2a the mean site effects for each value of m , averaged across the $B = 100$ subsamples, as a dotted line (Site 1 in blue, Site 2 in grey), along with estimated 95% confidence intervals (shaded areas). The true site effects estimated from the full dataset ($m = 210$) are represented by the horizontal solid lines. One can observe that site effects estimated from smaller samples approximate well the true site effects.

In Figure A.2b, we calculated the root mean square error (RMSE) between (1) ComBat-harmonized FA values using site effects estimated from the full dataset and (2) ComBat-harmonized FA values using site effects estimated from subsamples of size m . Again, the dotted line represent the average RMSE across the $B = 100$ subsamples, and the shaded area represents a estimated 95% confidence interval for the RMSE. For all values of m , the average RMSE is rather small and much smaller than the true site effects. Consistent with the results of Figure A.2a, the RMSE improves as a function of sample size.

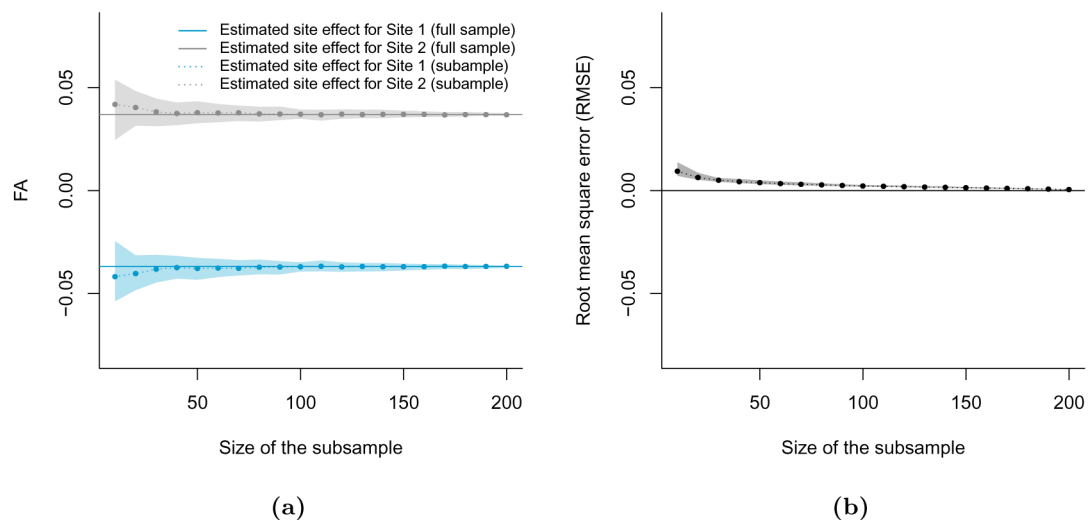


Figure A.2. Stability analysis of the ComBat harmonization

(a) The dotted lines represent the average site effects estimated by ComBat for each of the site, for each subsample size, averaged across the B subsamples. The shaded areas depict 95% confidence intervals. The solid lines represent the site effects estimated by ComBat on the full dataset ($m=210$). (b) The root mean square error (RMSE) between (1) ComBat-corrected FA values using site effects estimated on subsamples of size m and (2) The ComBat-corrected FA values using site effects estimated on the full sample ($m=210$).

Appendix B

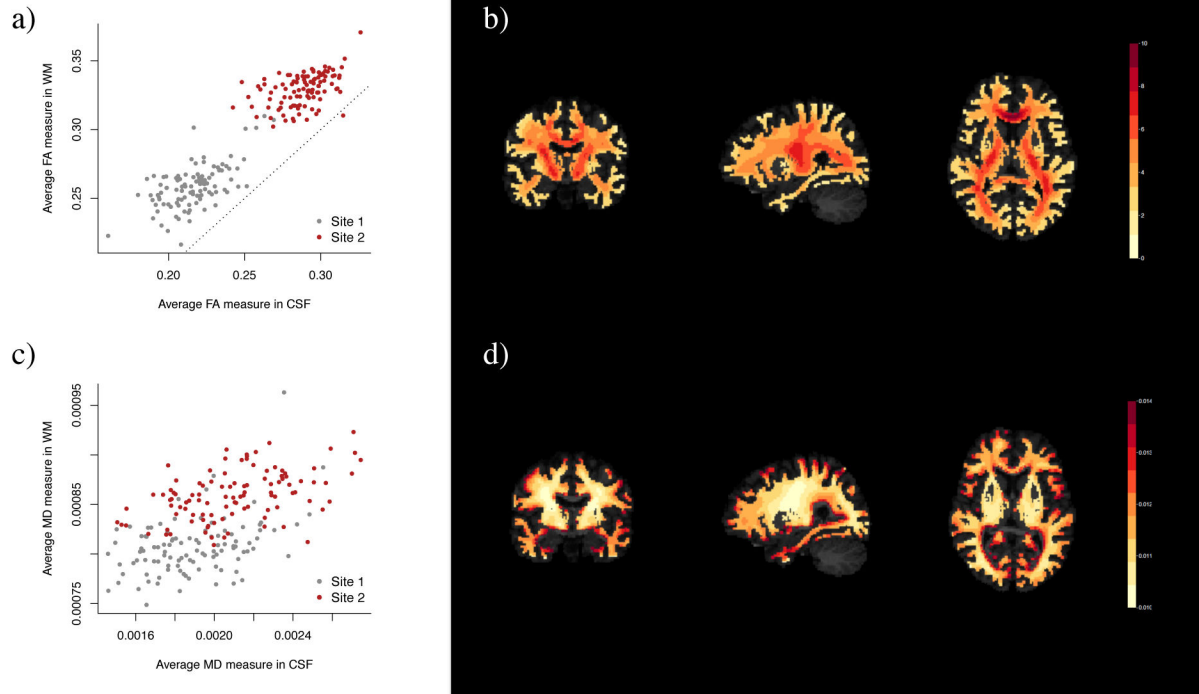


Figure B.1. RAVEL harmonization

(a) Relationship between the average FA measure in white matter (WM) and cerebrospinal fluid (CSF). The FA measurements vary by site in both WM and CSF. (b) Voxel-specific RAVEL coefficient $\hat{\psi}_V$ in template space for FA maps. (c) Relationship between the average MD measure in white matter (WM) and cerebrospinal fluid (CSF). The MD measurements vary by site in WM, but do not seem to vary in CSF. (d) Voxel-specific RAVEL coefficient $\hat{\psi}_V$ in template space for MD maps.

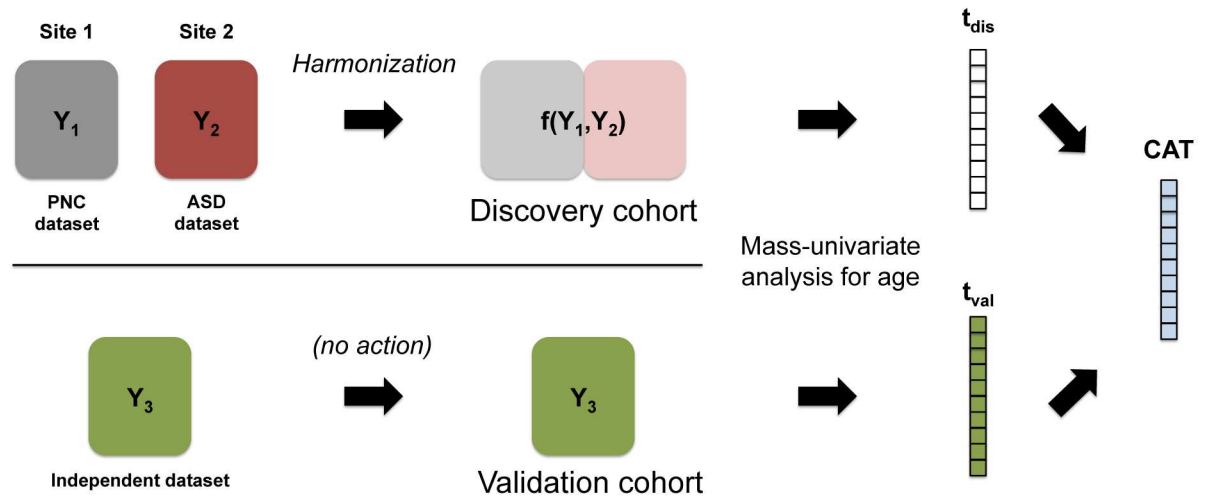


Figure B.2. Discovery-validation scheme for the estimation of replicability

To estimate the performance of a harmonization procedure at improving the replicability of the voxels associated with age, we use the harmonized dataset as a discovery cohort, and an independent dataset (different participants) as a validation cohort. For each cohort separately, we perform a mass-univariate analysis for age to obtain a t-statistic at each voxel. This yields two vectors of t-statistics, t_{dis} and t_{val} for the discovery and validation cohorts respectively. We calculate the agreement between t_{dis} and t_{val} using the concordance at the top (CAT) curve, described in the Methods section. A harmonization method that performs better will yield a vector t_{dis} more similar to t_{val} , that is a CAT curve closer to 1.

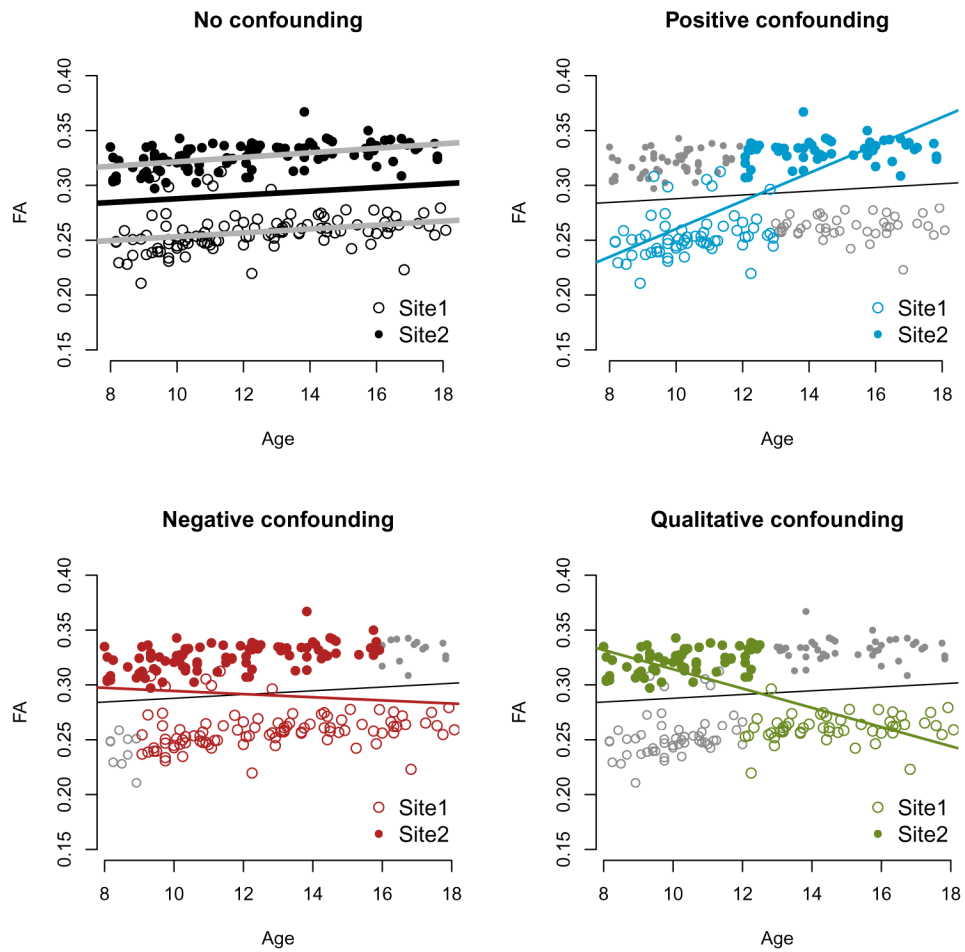


Figure B.3. Confounding scenarios for FA maps

Same as Figure 5, but for the per-scan median FA value in the White Matter (WM).

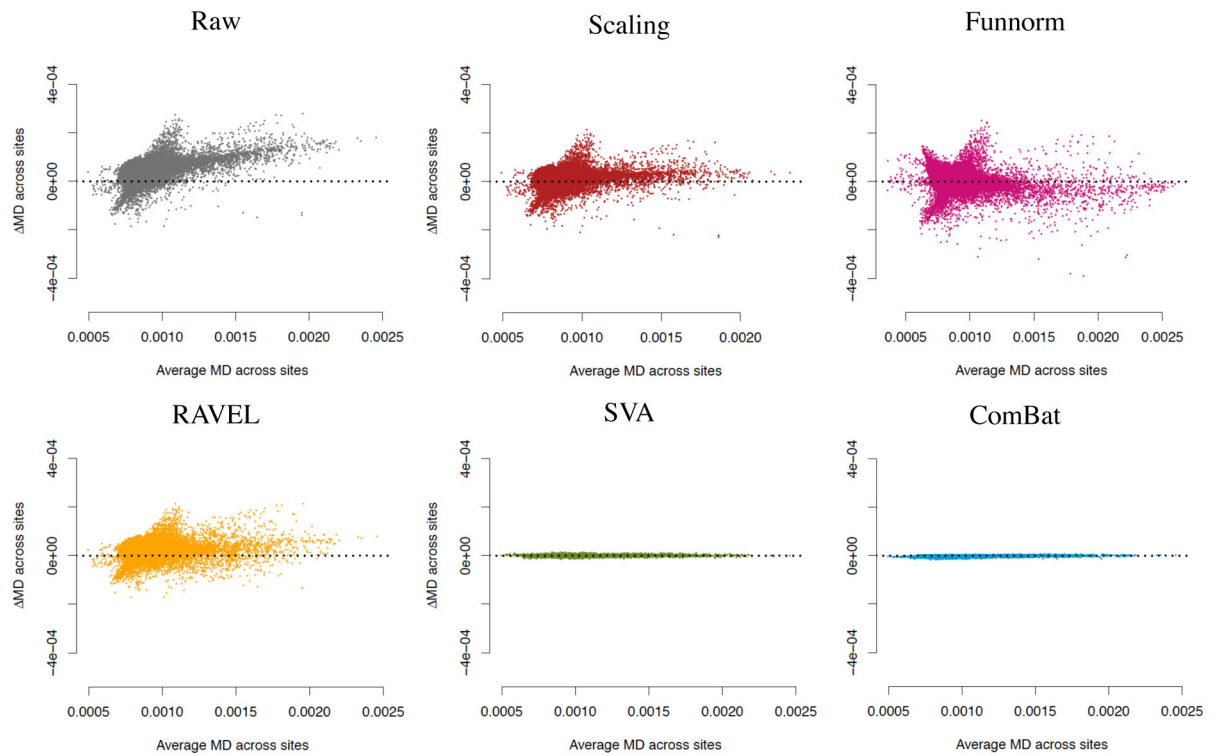


Figure B.4. MA-plots for site differences in MD maps
Same as Figure 3, but for MD maps.

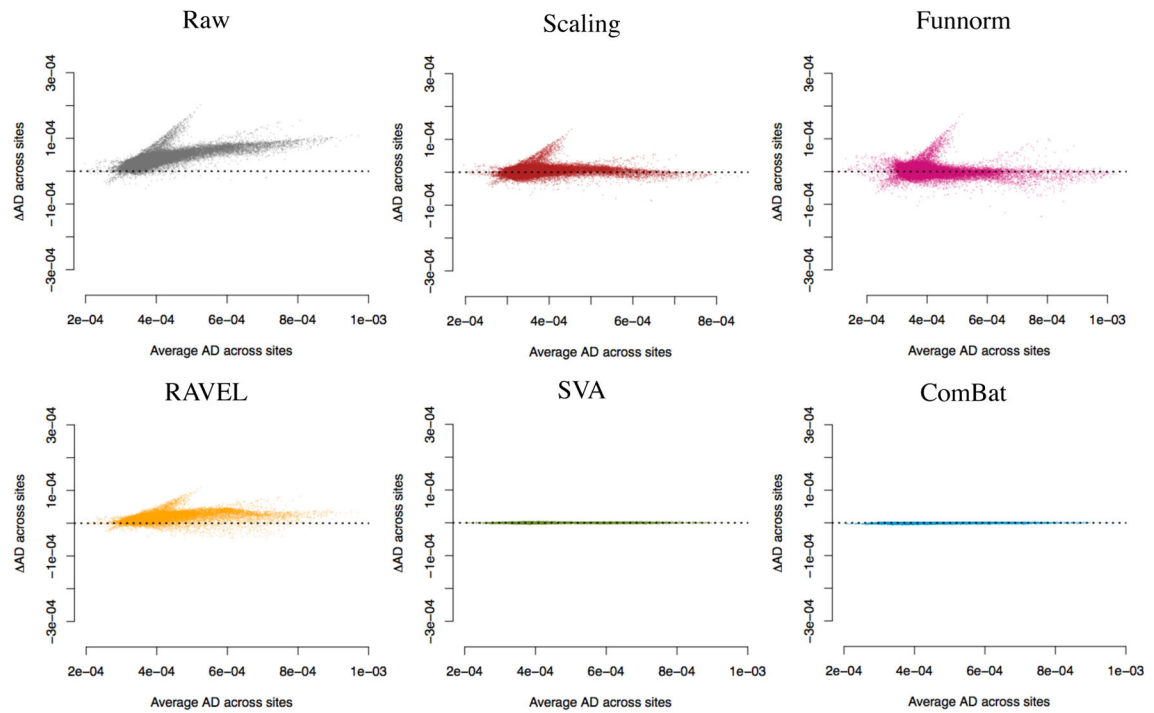


Figure B.5. MA-plots for site differences in AD maps

Same as Figure 3, but for AD maps.

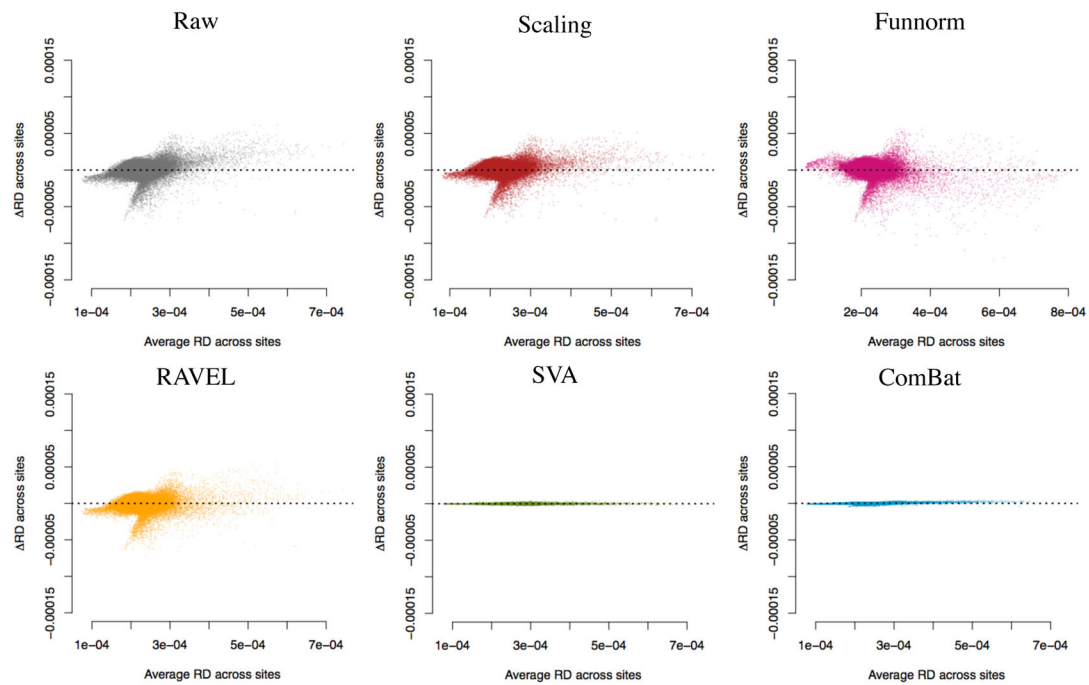


Figure B.6. MA-plots for site differences in RD maps

Same as Figure 3, but for RD maps.

a) Number of significant ROIs associated with site

b) Number of significant ROIs associated with age

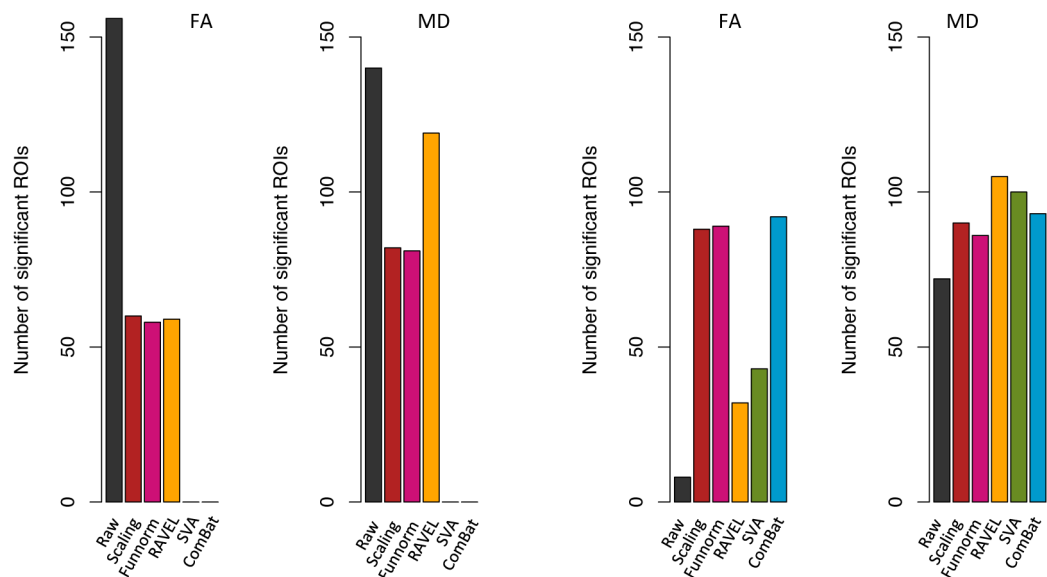


Figure B.7. Number of ROIs associated with site and age

Same as Figure 4, but for the 156 regions of interest (ROIs). All p-values were adjusted for multiple comparisons in a conservative manner using Bonferroni correction. (a) In the

absence of harmonization (raw data), all 156 ROIs are associated with site in the FA maps, and 140 ROIs are associated with site in the MD maps. Both SVA and ComBat result in 0 ROI associated with site. **(b)** ComBat performs well at increasing the number of ROIs associated with age (92 ROIs for FA and 92 ROIs for MD), as opposed to 8 ROIs and 72 ROIs in the raw data, for the FA and MD maps respectively.

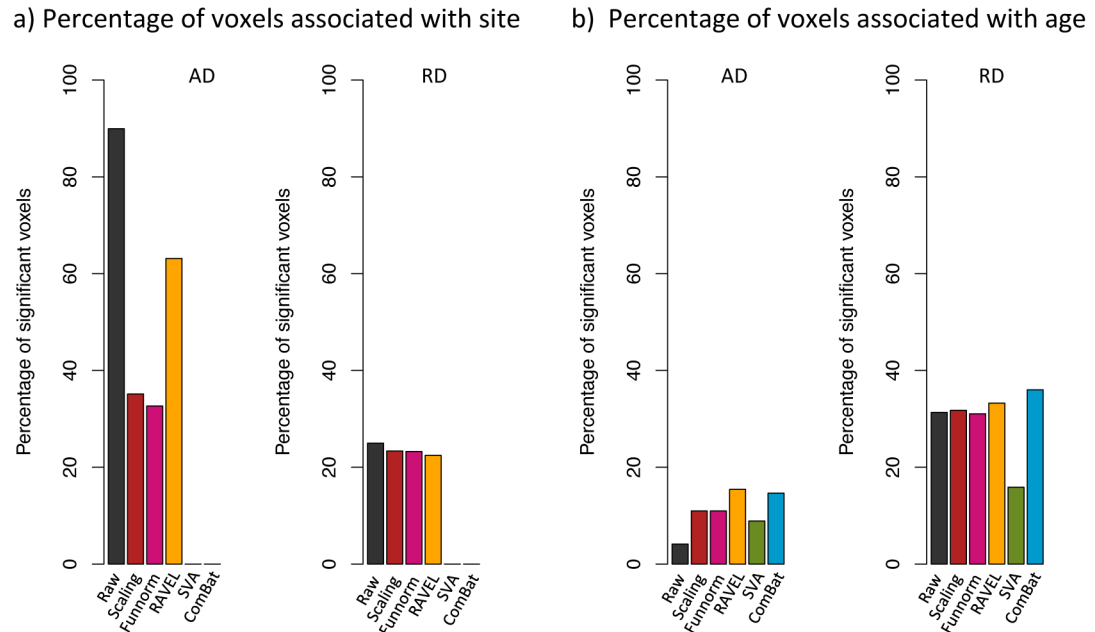


Figure B.8. Percentage of voxels associated with site and age for AD and RD maps
Same as Figure 4, but for the AD and RD maps.

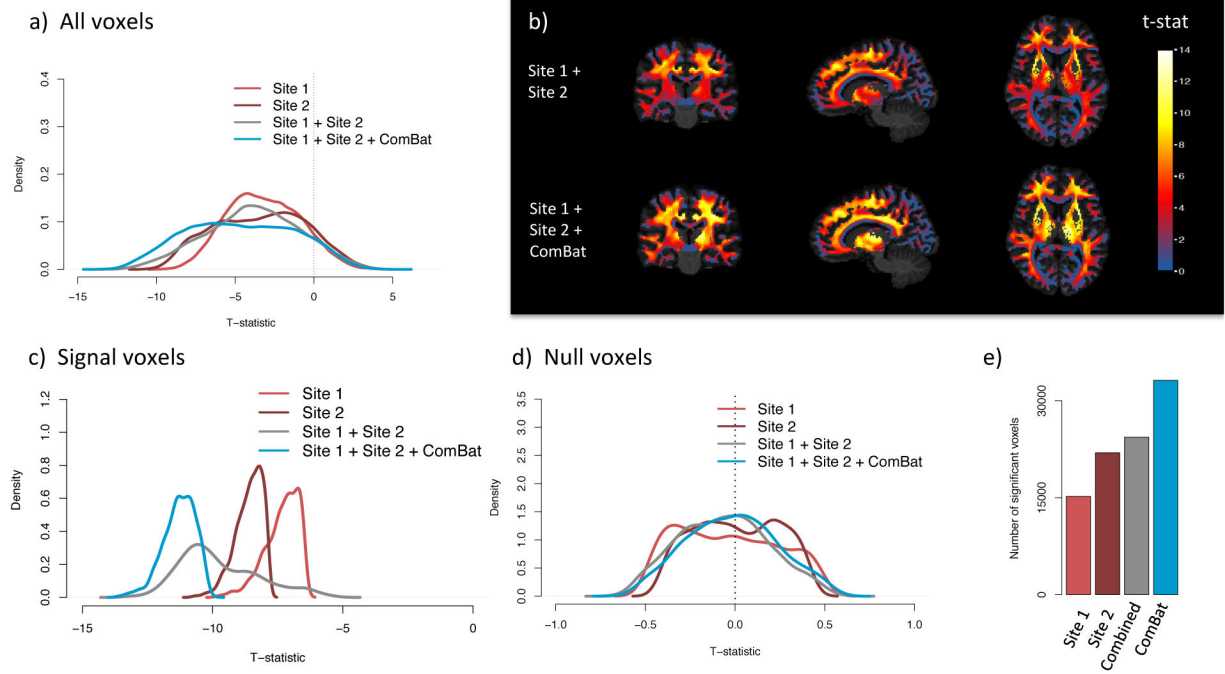


Figure B.9. Effect of ComBat harmonization on t-statistics (MD maps)
Same as Figure 8, but for the MD maps.

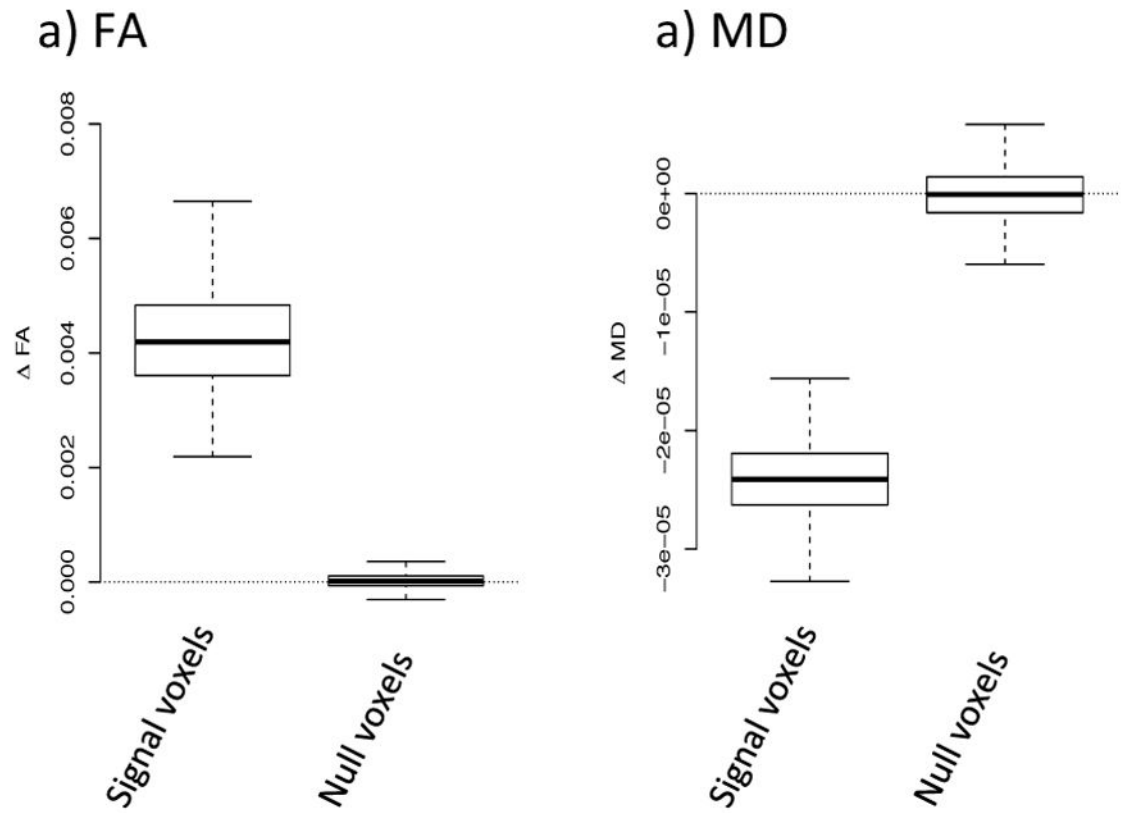
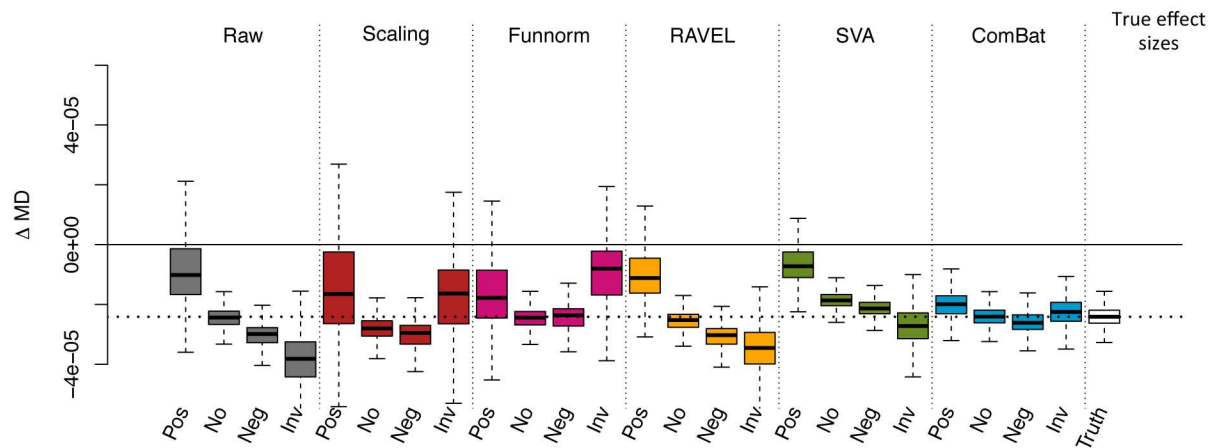


Figure B.10. Distribution of the effect sizes for the silver-standards

a) Signal voxels



b) Null voxels

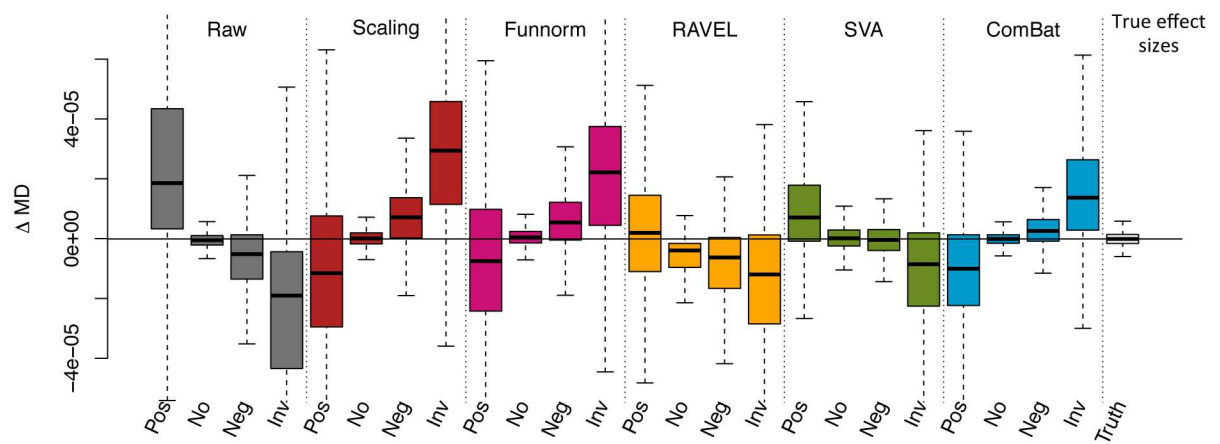


Figure B.11. Estimated effect sizes $\hat{\Delta}_{ageMD}$ for different confounding scenarios
Same as Figure 7, but for MD.

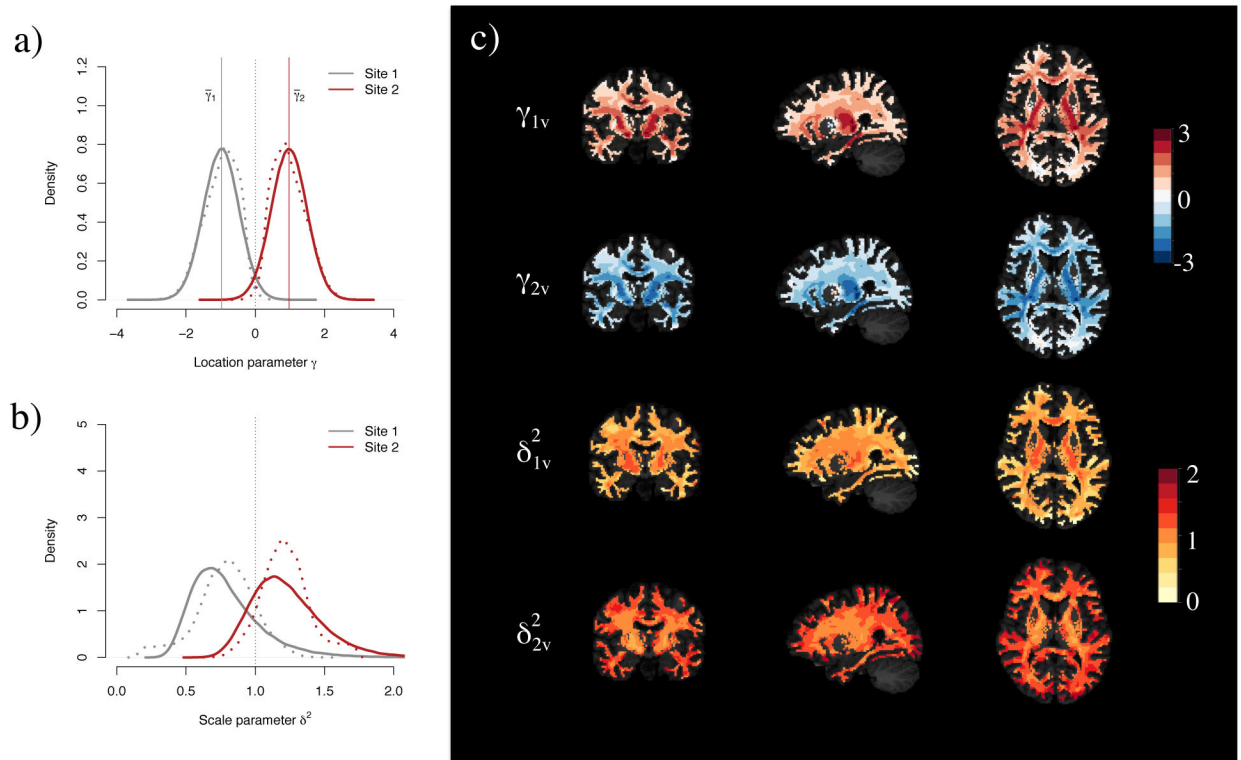


Figure 1. ComBat site effect parameters for FA

(a) The voxel-wise estimates of the location parameter γ_{iv} for site 1 (dotted grey line) and site 2 (dotted red line) for the FA maps. The solid lines represent the prior distributions (normal distributions with mean $\bar{\gamma}_1$ and $\bar{\gamma}_2$ respectively) estimated in the ComBat procedure using empirical Bayes. **(b)** The voxel-wise estimates of the scale parameter δ_{iv}^2 for site 1 (dotted grey line) and site 2 (dotted red line). The solid lines represent the EB-based prior distributions (inverse gamma distributions) estimated in the ComBat procedure. **(c)** Final EB-estimates for the site effects parameters for site 1 (first and third row) and site 2 (second and fourth row) in template space.

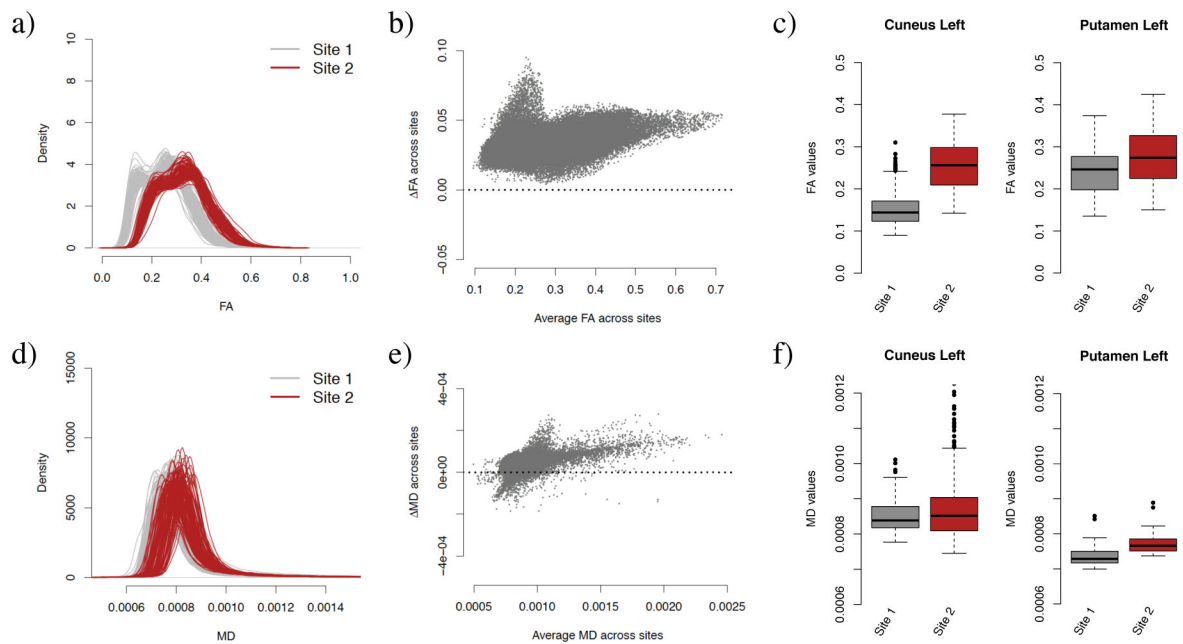


Figure 2. FA and MD maps are affected by site

(a) Density of the FA values for WM voxels for each participant, colored by site. (b) MA-plot for site differences in FA. The y-axis represents the differences in FA between Site 1 and Site 2, while the x-axis shows the average FA across sites. FA maps that would be free of site effects would result in an MA-plot centered around 0. The upper-left part of the scatterplot shows that several voxels appear to be differently affected by site in comparison to the rest of the voxels. (c) Boxplot of FA values for voxels located in two regions of interest (Cuneus left and Putamen left), depicted per site (FA values were averaged by site at each voxel separately). This shows that the magnitude of the difference in means between the two sites is region-specific. (d-f) Same as (a-c), but for the MD maps.

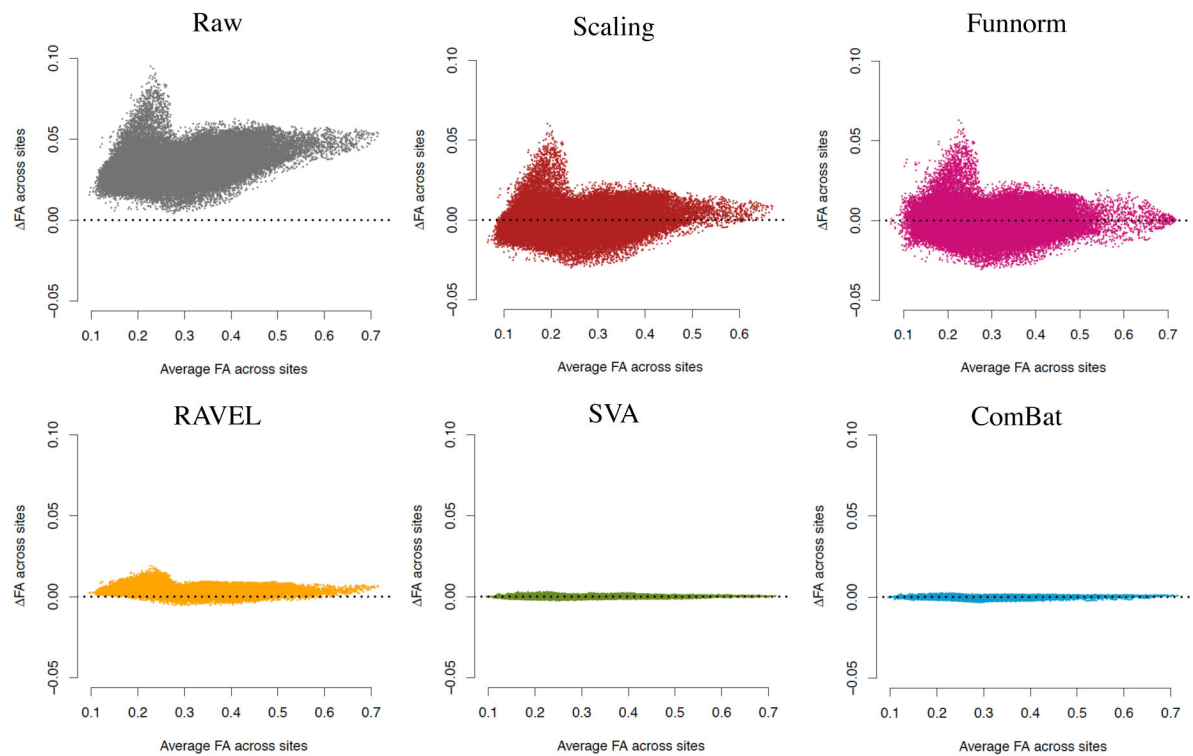


Figure 3. MA-plots for site differences in FA maps

Mean-difference (MA) plot for the FA maps for the different harmonization methods. At each voxel in the WM, the y-axis represents the difference between the average FA value at site 1 and the average FA value at site 2, and the x-axis represents the average FA value across all participants from both sites. A dataset free of site effects will result in MA data points near $y = 0$ for all values of x .

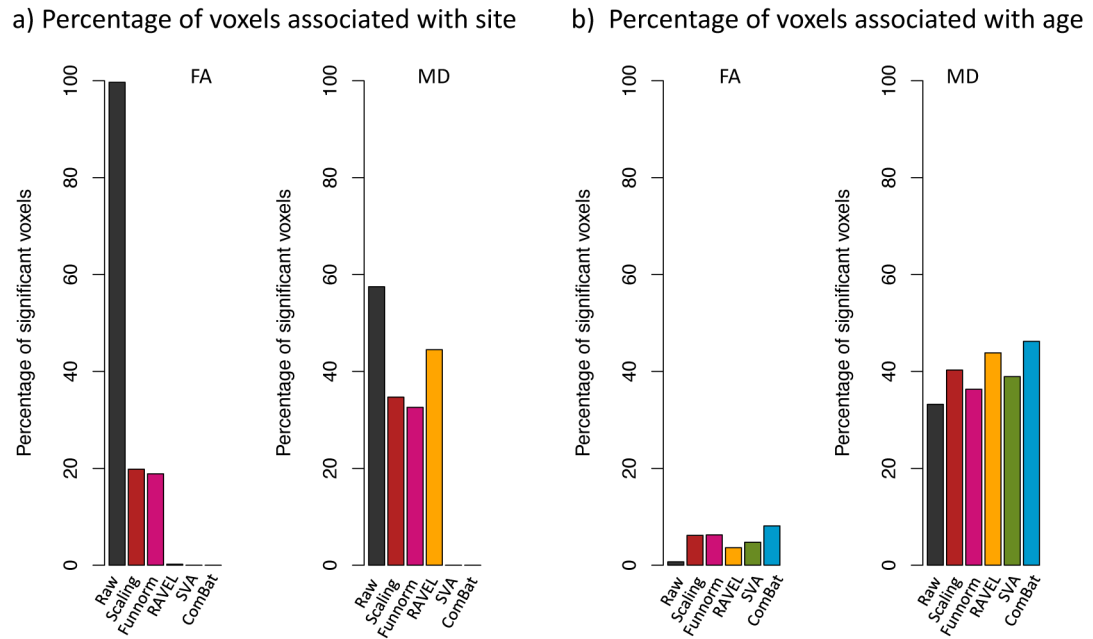


Figure 4. Percentage of voxels associated with site and age

(a) For each harmonization method, we calculated the number of voxels in the white matter (WM) that are significantly associated with imaging site for both FA and MD. A voxel is significant if the p -value calculated from a two-sample t -test is less than $p < 0.05$, after adjusting for multiple comparisons using Bonferroni correction. Lower numbers are desirable. (b) Number of voxels in the WM that are significantly associated with age using simple linear regression ($p < 0.05$) for both FA and MD. Higher numbers are desirable. From a total of 69,693 voxels in the WM, 69,475 and 40,056 voxels are associated with site in the raw data, for the FA and MD maps respectively. Both SVA and ComBat successfully remove the association with site for all voxels. ComBat performs the best at increasing the number of voxels associated with age (5,658 voxels for FA and 32,203 voxels for MD).

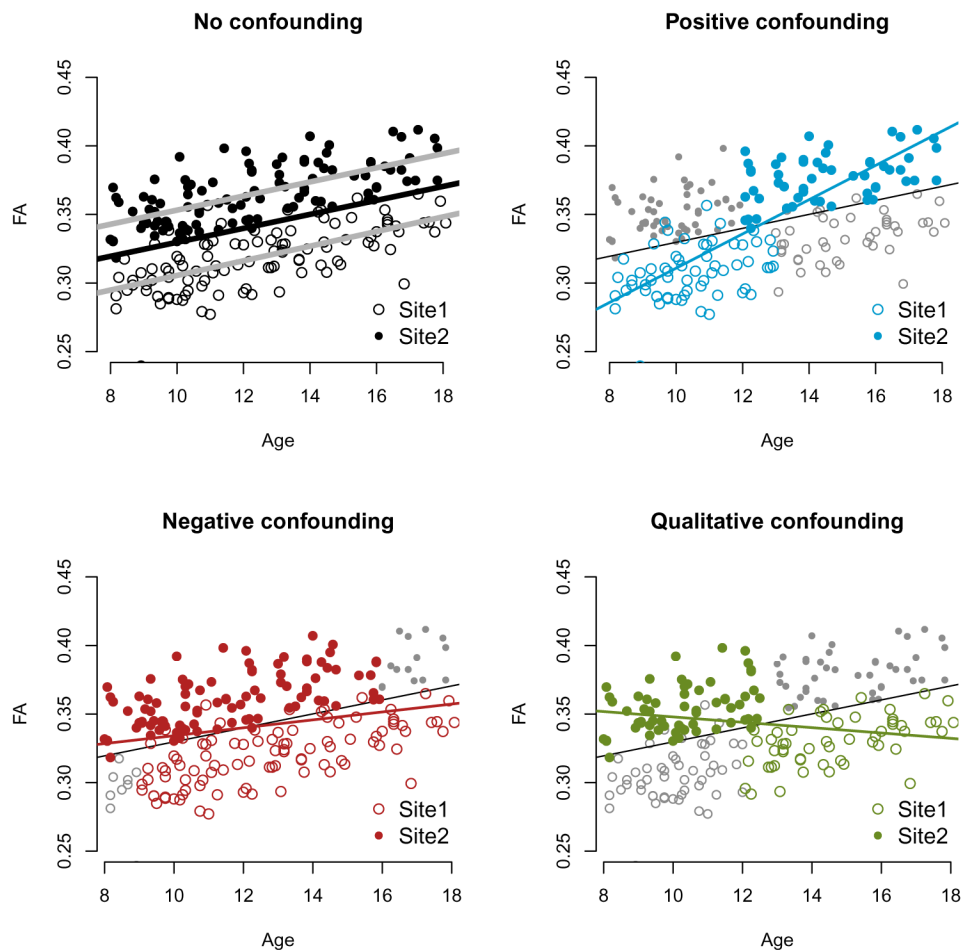


Figure 5. Confounding scenarios for FA maps

In all four panels, each data point represents the FA value versus the age of the participant for a fixed voxel in the right thalamus. Full dots and circles are used to distinguish the two sites of the participant scans (Dataset 1 and Dataset 2). The solid black line in all panels represents the estimated linear relationship between FA and age when all data points are included (absence of confounding). In panel (a), the grey lines represent the estimated relationship between FA and age for each site. In panels (b–d), the selected participants are colored (blue, red and green respectively), and the colored solid lines represent the estimated linear relationship between FA and age for the selected participants only.

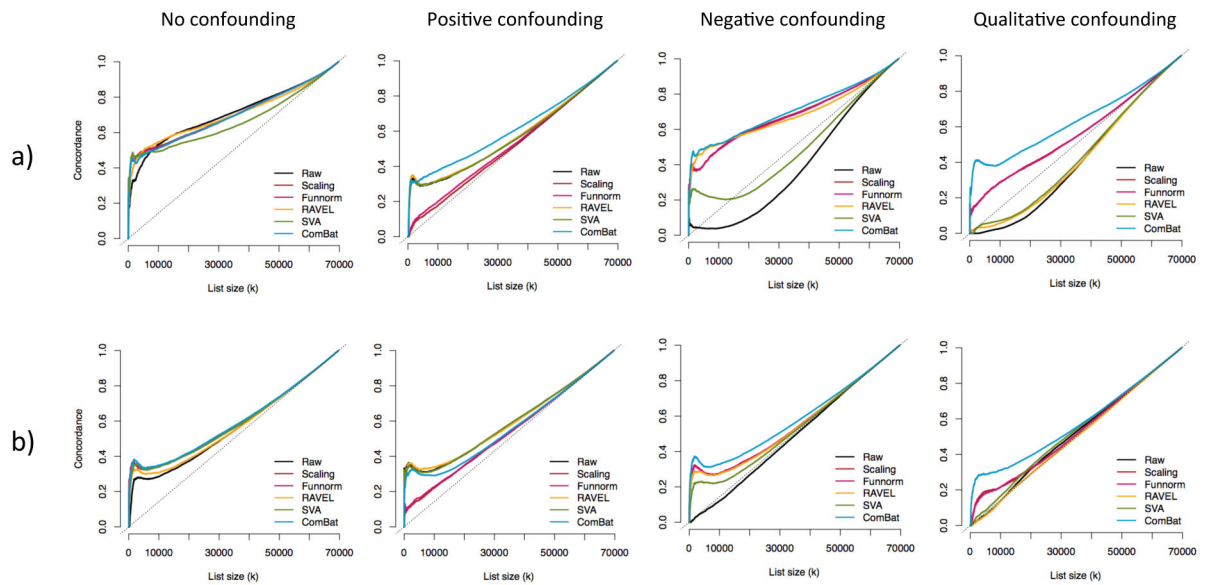


Figure 6. Replicability of the voxels associated with age in the FA maps

For each confounding scenario and for each harmonization method, we calculated a concordance at the top (CAT) curve for the voxels associated with age. The concordances were calculated between the harmonized dataset (2 sites combined) and an independent dataset. In **(a)**, 292 unrelated participants within the same range were selected as an independent cohort. In **(b)**, 105 unrelated and older participants were selected as an independent cohort. A good harmonization will result in a CAT curve closer to 1. Overlaps by chance will result in a CAT curve along the diagonal.

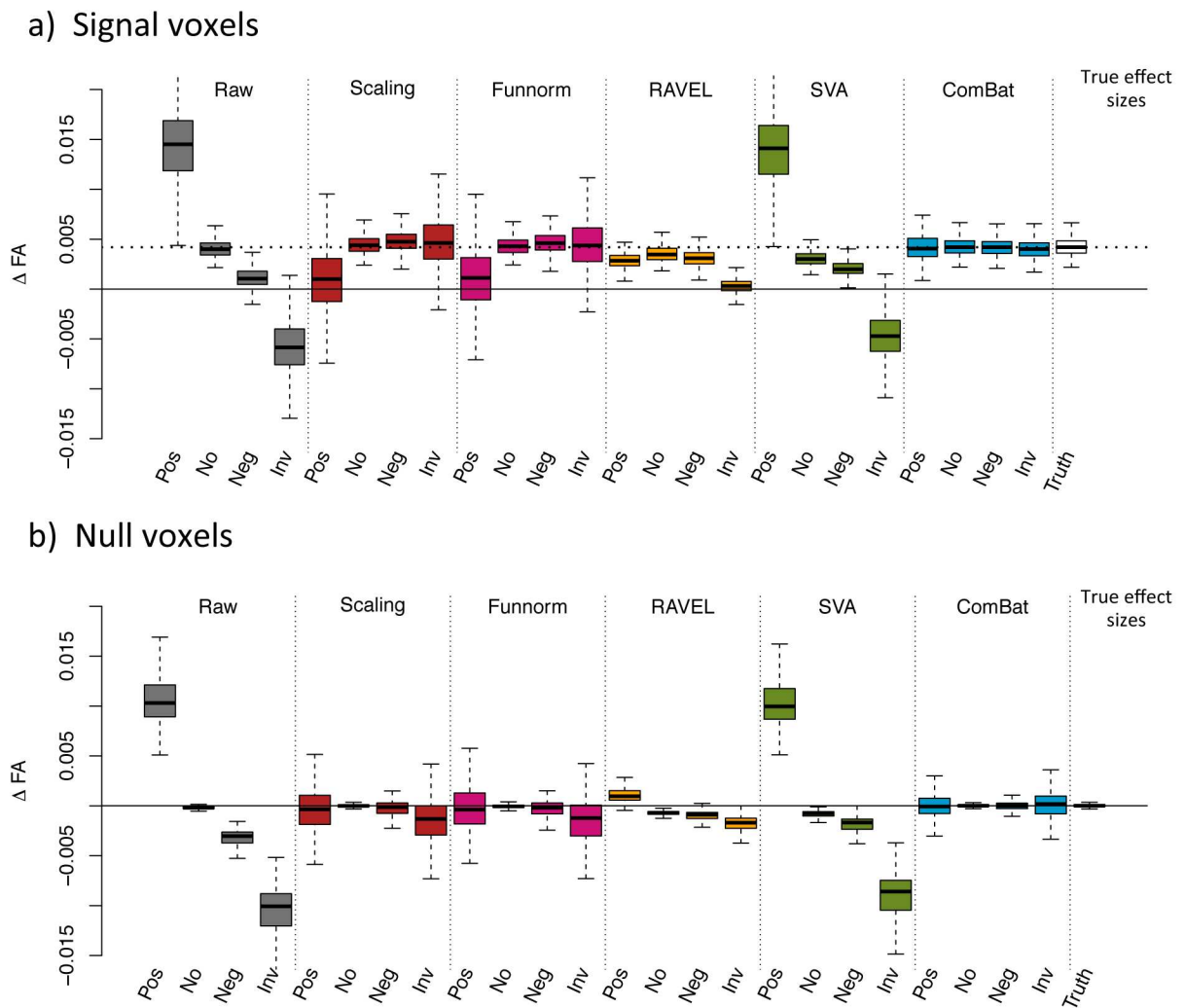


Figure 7. Estimated effect sizes $\hat{\Delta}_{age}FA$ for different confounding scenarios

(a) Boxplots of the estimated effect sizes $\hat{\Delta}_{age}FA$ for the set of signal voxels described in Section 3.8, for different confounding scenarios: positive confounding (pos), no confounding (no), negative confounding (neg) and quantitative confounding (rev). The dotted line represents the median true effect size (around 0.004). **(b)** Boxplots of the estimated effect sizes $\hat{\Delta}_{age}FA$ for the set of null voxels described in Section 3.8. The median true effect size is around 0. The distributions of the estimated effect sizes for the ComBat-harmonized datasets approximate very well the distribution of the true effect sizes shown in the last column in each panel. Results for MD values are presented in Figure B.11.

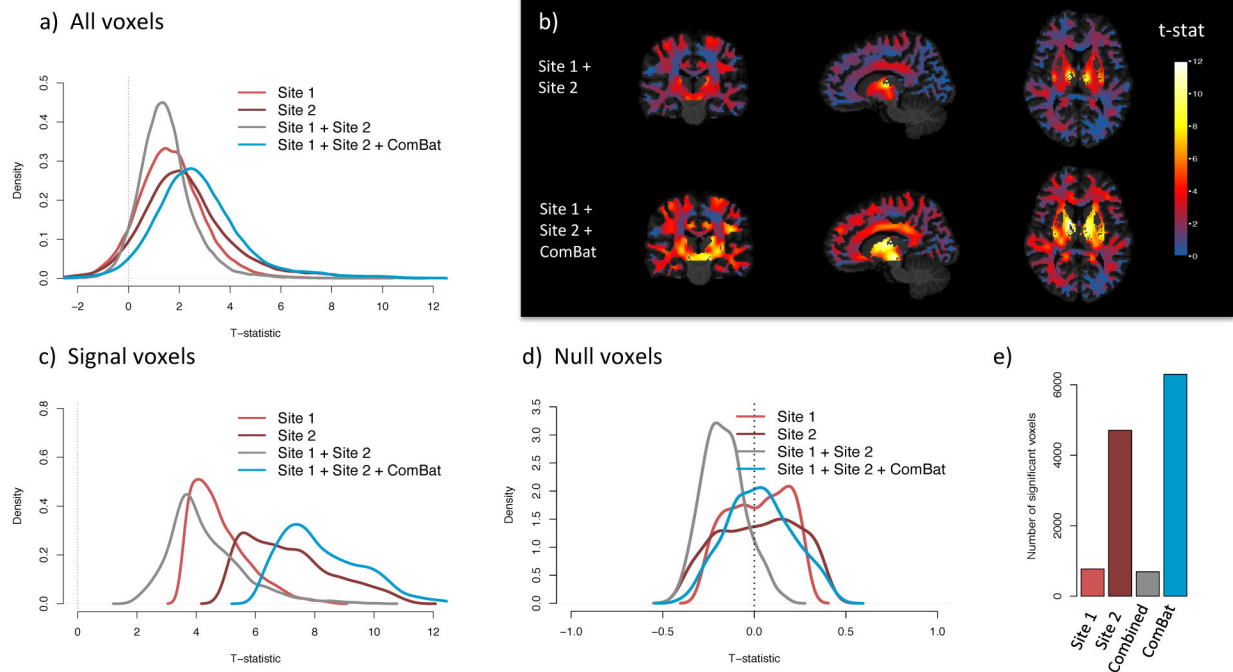


Figure 8. ComBat improves statistical power

We present voxel-wise t-statistics in the WM, testing for association between FA values and age, for four combinations of the data: Dataset 1 and Dataset 2 analyzed separately, Dataset 1 and Dataset 2 combined without any harmonization, and Dataset 1 and Dataset 2 combined and harmonized with ComBat. **(a)** Distribution of the t-statistics for all WM voxels, for each analyzed dataset. The combined datasets harmonized with ComBat show higher t-statistics. **(b)** T-statistics in template space for the combined dataset, with no harmonization (top row) and with ComBat (bottom row). **(c)** Distribution of the t-statistics for a subset of voxels highly associated with age (signal silver-standard described in Section 2.5). **(d)** Distribution of the t-statistics for a set of voxels not associated with age (null silver-standard described in Section 2.5). ComBat increases the magnitude of the t-statistics for the signal voxels while maintaining the t-statistics around 0 for the null voxels. **(e)** Number of voxels significantly associated with age. Bonferroni correction was applied to correct for multiple comparisons.

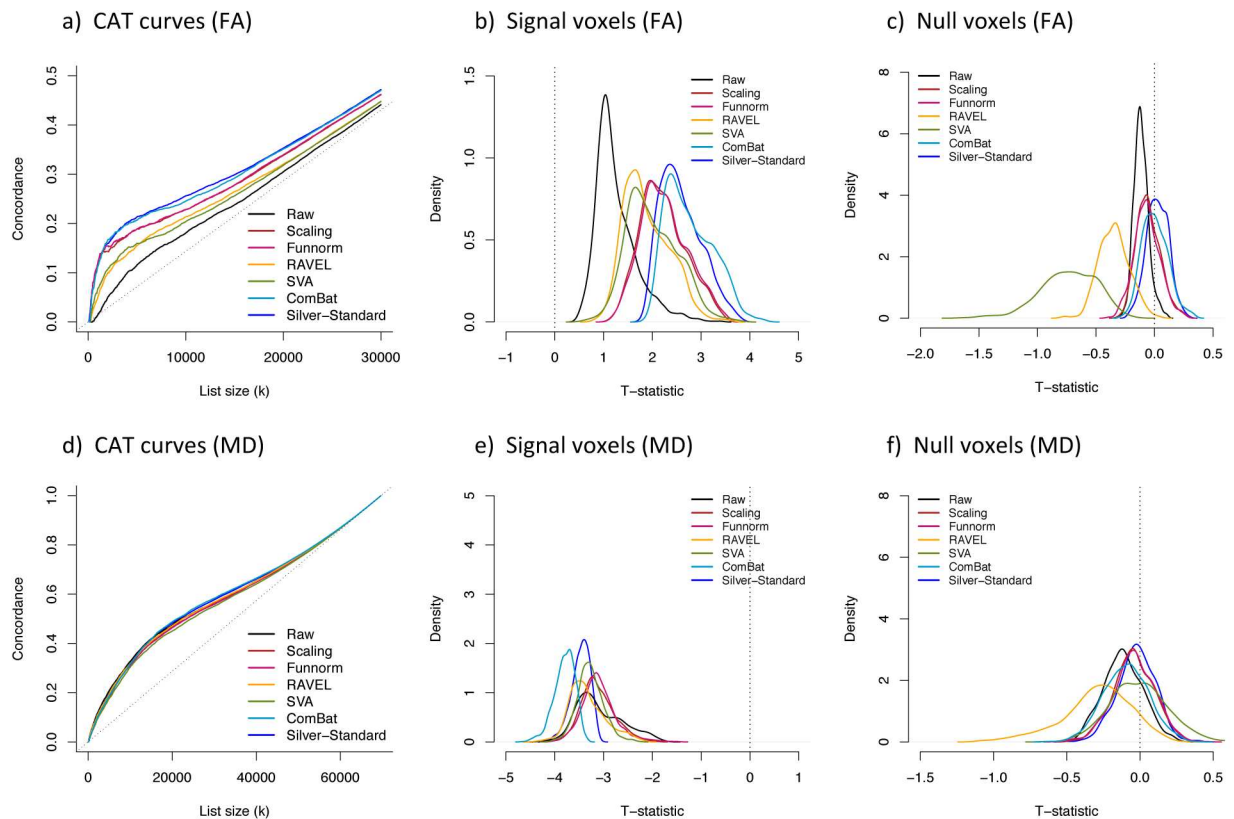


Figure 9. ComBat is robust to small sample size studies

We created $B = 100$ random subsets of size 20, selecting at random 10 participants from each site, and applied each harmonization method on every subset separately. For each harmonized subset, we computed a t-statistic at each voxel in the WM, testing for the association of FA and MD with age. We created a silver-standard list of t-statistics by creating $B = 100$ random subsets of size 20 within site. **(a)** Average concordance at the top (CAT) curve for each harmonization method for the FA maps. The silver-standard CAT curve is depicted in dark blue. A higher curve represents better replicability of the voxels associated with age. **(b)** Densities of the t-statistics for the set of signal voxels described in Section 3.8, for the FA maps. Higher values of the t-statistics are desirable. **(c)** Densities of the t-statistics for the set of null voxels described in Section 3.8, for the FA maps. T-statistics closer to 0 are desirable. For each plot, the results obtained for the ComBat-harmonized datasets approximate very well the results obtained from the within-site silver-standard (dark blue). **(d)** Same as (a), but for the MD maps. **(e)** Same as (b), but for the MD maps. Lower values of the t-statistics are desirable. **(f)** Same as (c), but for the MD maps. RAVEL performs substantially worse than other methods.