NeuroImage xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Variance decomposition for single-subject task-based fMRI activity estimates across many sessions

Javier Gonzalez-Castillo[a,*], Gang Chen[b], Thomas E. Nichols[c], Peter A. Bandettini[a,d]

[a] Section on Functional Imaging Methods, Laboratory of Brain and Cognition, National Institute of Mental Health, Bethesda, MD, United States
[b] Scientific and Statistical Computing Core, National Institute of Mental Health, Bethesda, National Institutes of Health, Bethesda, MD, United States
[c] Department of Statistics & WMG, University of Warwick, Coventry, UK
[d] Functional MRI Facility, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States

ARTICLE INFO

Keywords:
Task-based fMRI
Variance decomposition
Longitudinal studies

ABSTRACT

Here we report an exploratory within-subject variance decomposition analysis conducted on a task-based fMRI dataset with an unusually large number of repeated measures (i.e., 500 trials in each of three different subjects) distributed across 100 functional scans and 9 to 10 different sessions. Within-subject variance was segregated into four primary components: variance across-sessions, variance across-runs within a session, variance across-blocks within a run, and residual measurement/modeling error. Our results reveal inhomogeneous and distinct spatial distributions of these variance components across significantly active voxels in grey matter. Measurement error is dominant across the whole brain. Detailed evaluation of the remaining three components shows that across-session variance is the second largest contributor to total variance in occipital cortex, while across-runs variance is the second dominant source for the rest of the brain. Network-specific analysis revealed that across-block variance contributes more to total variance in higher-order cognitive networks than in somatosensory cortex. Moreover, in some higher-order cognitive networks across-block variance can exceed across-session variance. These results help us better understand the temporal (i.e., across blocks, runs and sessions) and spatial distributions (i.e., across different networks) of within-subject natural variability in estimates of task responses in fMRI. They also suggest that different brain regions will show different natural levels of test-retest reliability even in the absence of residual artifacts and sufficiently high contrast-to-noise measurements. Further confirmation with a larger sample of subjects and other tasks is necessary to ensure generality of these results.

Introduction

Functional MRI (fMRI) time series constitute high-dimensional, rich spatio-temporal recordings of brain function that can be modulated by different physiological (e.g., anxiety levels), neuronal (e.g., ongoing cognition) and experimental factors (e.g., time-of-the-day) surrounding a scanning session. Activation and connectivity fMRI maps are not only dependent on the amount of residual head motion (Power et al., 2012), physiological noise (Birn, 2012) and hardware instabilities (Jo et al., 2010) not properly accounted for during pre-processing; but also vary as a function of additional factors such as attention (Vuilleumier and Driver, 2007), learning (Dayan and Cohen, 2011), caffeine ingestion (Liu et al., 2004), sleep (Gaggioni et al., 2014; McKenna et al., 2014), metabolite concentrations in blood (Poldrack et al., 2015) and, potentially, gene expression levels (Poldrack et al., 2015). Experimenters cannot always control for all these factors, which end up adding unexplained within-subject variance to the data, and obstructing interpretation of single-subject longitudinal results.

Additionally, signal fluctuations of interest in fMRI (i.e., those of a BOLD origin and driven by underlying neuronal activity) only account for a small percentage of the variance present in the data (Bianciardi et al., 2009). As a result of this, fMRI has been traditionally regarded as a technique with limited sensitivity due to insufficient contrast-to-noise ratio (CNR). This is particularly true within the context of potential clinical applications. While group averaging can alleviate insufficient CNR in a research environment, combining data across subjects is not an option in a clinical setting. Alternatively, single-subject CNR can be improved by combining successive within-subject recordings as long as the signal of interest remains relatively constant and the noise is randomly distributed across those repeated measures. In fact, intra-subject trial averaging is a common practice in other neuroimaging modalities such as in electroencephalography. For example, over a thousand trials are routinely combined to reliably detect brainstem auditory evoked-response potentials (ERPs) (Skoe and Kraus, 2010),

* Corresponding author.

http://dx.doi.org/10.1016/j.neuroimage.2016.10.024
Received 11 July 2016; Accepted 14 October 2016
Available online xxxx

and several hundred are combined when aiming for cortical ERPs in occipital cortex (e.g., visual P1 waves), where CNR is much higher (Luck, 2014). Although obtaining such "*high-N*" in individual subjects is not a common practice in fMRI, a few recent studies have demonstrated that when doing so (N_{runs}≈100) a richer, highly distributed picture of brain function emerges (Gonzalez-Castillo et al., 2012, 2015). Moreover, when those within-subject "*high-N*" experiments are accompanied by intensive phenome-wide assessments, the joint dynamics of human brain and metabolic function can be assessed in detail (Poldrack et al., 2015). Yet, collecting hundreds of trials in task-based fMRI may require several sessions, which in turn adds an additional component to total within-subject variance. Given the above-mentioned benefits associated with acquiring *"high-N"* within-subject measures, and the importance of within-subject longitudinal studies for developmental and clinical research, a better understanding of how within-subject variance decomposes across basic experimental units (e.g., runs, sessions) is desirable, so that multi-session fMRI experiments can be optimized to minimize within-subject variance.

Although substantial past efforts have been devoted to assess the test-retest reliability of task-based fMRI(Gonzalez-Castillo and Talavage, 2011; Gountouna et al., 2010; Havel et al., 2006; McGonigle, 2012; McGonigle et al., 2002; Plichta et al., 2012), most of them rely on a limited number of sessions ($N_{session} < 5$). This is not sufficient to attempt any decomposition of within-subject variance into its primary subcomponents: (a) across-sessions variance ($\sigma^2_{session}$; i.e., that associated with entering and exiting the scanner on the same or different days); (b) across-runs variance (σ^2_{run}; with run defined as a continuous scanning period that contains several blocks/trials of stimulation and/or task); (c) across-blocks variance (σ^2_{block}; with block defined as each individual contiguous occurrence of the task/stimulus); and (d) error/modeling variance (σ^2_{err} ; i.e., the remaining within-subject variance not attributable to any of the other three factors described here). Here, we address this gap by performing an exploratory variance-decomposition analysis in one of the within-subject *"high-N"* datasets mentioned above (Gonzalez-Castillo et al., 2012) using a four-level nested random-effect variance decomposition model. The selected dataset is particularly well suited for this exploration because it contains fMRI recordings for over 500 trials (a trial here is a 60 s second block with 20 s of task and 40 s of rest; please see details below) in each of three individual subjects. Those 500 trials were all acquired under the same experimental condition (i.e., a visual stimulation plus a letter/number discrimination task), and across 100 functional runs distributed among 9 to 10 different scanning sessions over a 3-month period.

Given this larger-than-usual number of recordings, we were able to segregate variance in the four components cited above (i.e., $\sigma^2_{session}$, σ^2_{run}, σ^2_{block} and σ^2_{error}). Our results not only show how measurement/modeling variance dominates over the whole brain, but more interestingly, how the other three components show a non-homogenous spatial distribution that is reproducible across subjects. In particular, $\sigma^2_{session}$ is the second dominant source of within-subject variance in occipital cortex (i.e., the primary input region for the task), while σ^2_{run} is in most other regions. A detailed evaluation across sixteen well established cortical networks (Laird et al., 2011) revealed how σ^2_{block} contributes to within-subject variance to a much larger extent in higher-order cognitive networks than in somatosensory networks. In fact, for a subset of higher-order cognitive networks (specifically those previously associated with emotion/interoception roles) a large percentage of voxels show $\sigma^2_{block} > \sigma^2_{session}$.

These novel insights into the spatio-temporal distribution of within-subject variance, not only confirm previous accounts of the overwhelming contribution of modeling/measurement error to within-subject variance (Friedman et al., 2006; Suckling et al., 2008), but could also help optimize future multi-session single-subject studies. For example, the dominance of across-session over across-run variance in visual regions suggests having fewer longer sessions rather than many shorter sessions, this if solely interested in responses within occipital cortex. The dominance of across-run over across-session variance everywhere else in the brain suggests that limiting sessions should not be a primary optimization criteria if interested in evaluating responses beyond occipital cortex.

The code necessary to perform the four-level variance decomposition will be made publicly available upon publication as part of the AFNI software suite. This software, if directly provided with run-level activity estimates, could be used in group studies to decompose total group variance into variance across-subjects, across-sessions, and across-runs; and in that manner better segregate true between-subject variance in the experiments. In addition, the dataset presented here is also publicly available upon request in Xnat Central (https://central.xnat.org) under project ID: 100RunsPerSubj.

Methods

The analyses presented here were conducted on a task-based dataset previously described in (Gonzalez-Castillo et al., 2012) that contains a total of 100 runs acquired over 9 to 10 different sessions (on average 10.3 ± 2.4 runs per session) in each of three different individuals (one male/two females:age=27 ± 2.5 y.o.). Below we provide a brief description of the task and acquisition parameters. Please refer to the supplementary materials of the original study for additional details.

All participants gave informed consent in compliance with a protocol approved by the Institutional Review Board of the National Institute of Mental Health in Bethesda, MD.

Experimental task

All functional runs had the same organization of blocks. An initial 30 s rest period was followed by five repetitions of the following sequence of blocks: task block (20 s) and rest block (40 s). An additional 10 s of rest were added at the end of each functional run. This resulted in 340 s runs. During the rest periods, subjects were instructed to remain still and focus their attention on a white fixation cross over a black background. During the task epochs, subjects were instructed to focus their attention in the center of a flickering checkerboard (frequency=7.5 Hz) and to perform a letter/number discrimination task. Four random alpha-numeric characters appeared for 400 ms at random intervals in the center of the flickering checkerboard. Subjects were provided with a four-button response box (Curdes Fiber Optic Response Box Model No: HH-2x4-C) in their right hand and were instructed to press the leftmost button for each letter appearance and the next button if the character on the screen was a number.

Data acquisition

Imaging was performed on a General Electric (GE) 3 Tesla Signa HDx MRI scanner. Functional runs were obtained using a gradient recalled, single shot, full k-space echo planar imaging (geEPI) sequence [TR=2.0 s, TE=30 ms, FA=75°, 32 oblique slices, slice thickness=3.8 mm, spacing=0 mm, in-plane resolution=3.75×3.75 mm, field-of-view (FOV)=24 cm]. T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) sequence was also acquired for presentation and alignment purposes. Physiological data were recorded during functional runs using a pneumatic belt and an optical finger pulse oximeter. Acquisition of the dataset presented in this article required 10 visits for two subjects, and only 9 visits for the other subject. These visits spanned a period of around 3 months.

Data preprocessing

The Analysis of Functional NeuroImages (AFNI) software (Cox, 1996) was used for all of the data preprocessing. Preprocessing on each

individual EPI run included: (i) discard initial five volumes to allow for magnetization to reach steady-state; (ii) physiological noise removal using regressors that model the effects of respiration and cardiac cycle [RETROICOR (Glover et al., 2000)] as well as the effects of slow blood-oxygenation level fluctuations [RVT (Birn et al., 2008)]; (iii) slice-timing correction; (iv) intra-run motion correction; (v) within-subject inter-run spatial co-registration; (vi) spatial smoothing (FWHM=6mm); and (vii) intensity normalization, by dividing each voxel-wise time series by its own mean. Physiological noise removal was omitted for two runs in subject 1 because physiological data were not available.

Statistical analysis

Statistical analyses were performed separately in each subject after temporally concatenating all available 100 runs. We used AFNI program *3dREMLfit*, which accounts for temporal autocorrelation in the residuals of functional MRI (fMRI) time series using an ARMA (1, 1) model. Expected hemodynamic responses were modeled via convolving a gamma-variate function with a boxcar function that follows the experimental paradigm (e.g., "ones" during active blocks and "zeros" during rest/fixation blocks). This corresponds to the sustained-only model described in (Gonzalez-Castillo et al., 2012). Estimates of effect size (β) and associated *T-stat* were obtained for each separate task epoch (i.e., block). Nuisance regressors include run-specific 3rd order Legendre polynomials to account for slow drifts, and estimates of head motion and their first derivatives. This led to 500 estimates of effect size and their *T-stat* per subject that were input to the variance decomposition analysis described below.

Variance decomposition

Here we consider a model that partitions the total variance into four components that correspond to the following four hierarchical levels: within block (σ^2_{err}) and across blocks (σ^2_{block}), runs (σ^2_{runs}), and sessions ($\sigma^2_{\text{session}}$). We first start with a simple model, decomposing the effect estimate $\widehat{\beta}_{i(j(k))}$with the assumption of no measurement error,

$$\widehat{\beta}_{i(j(k))}=\alpha+\theta_k+\zeta_{j(k)}+\eta_{i(j(k))} \quad (1)$$

where indices *i, j, and k* denote the levels of block, run, and day, respectively; parentheses indicates the nesting structure between consecutive levels; α represents the intercept or overall average effect; θ_k, $\zeta_{j(k)}$, and $\eta_{i(j(k))}$ denote the session-, run-, and block-specific random effect, respectively, and are assumed to follow Gaussian distributions with a mean of zero and variances of σ^2_{block} , σ^2_{runs} and $\sigma^2_{\text{session}}$, respectively.

The framework (as in Eq. (1)) is basically a linear mixed-effects model with a sequentially nested random-effects structure, and the variance partition is straightforward,

$$var\left(\widehat{\beta}_{i(j(k))}\right)=\sigma^2_{session}+\sigma^2_{run}+\sigma^2_{block} \quad (2)$$

However, the model described in Eq. (1), with the assumption of no sampling error, is unrealistic because $\widehat{\beta}_{i(j(k))}$ is only an estimate of the ideal $\beta_{i(j(k))}$, the measurement not corrupted by measurement noise. Fortunately, this fourth source of variation is easily estimated through the regression analysis using a fixed-shaped hemodynamic response function (i.e., a canonical gamma-variate function). Therefore, we instead consider a more realistic model,

$$\widehat{\beta}_{i(j(k))}=\alpha+\varepsilon_k+\theta_{j(k)}+\eta_{i(j(k))}+\epsilon_{i(j(k))} \quad (3)$$

where $\epsilon_{i(j(k))}$ represents the measurement error that is assumed to follow a Gaussian distribution $G(0,\sigma_i^2)$, and $\sigma_i^2$is the estimated variance for the *ith* block. The variance composition for the model in Eq. (3) is then updated to,

$$var\left(\widehat{\beta}_{i(j(k))}\right)=\sigma^2_{session}+\sigma^2_{run}+\sigma^2_{block}+\sigma^2_{err} \quad (4)$$

The difference between the two models (Eq. (1)) and (Eq. (3)) can be conceptualized from a different perspective. Even with the presence of sampling errors, we could still work with the first model (Eq. (1)); however, the σ^2_{block}component in Eq. (2) would not really be the cross-block variance, but roughly the sum of the cross-block variance (σ^2_{block}) and the average (across blocks; σ^2_{err}) of the individual within-block variances σ_i^2 from Eq. (4). In other words, if all the effect estimates are equally reliable (i.e., have the same sampling variance), the component σ^2_{block} in Eq. (2) contains both the cross-block variance σ^2_{block} from Eq. (4) and the within-block sampling variance (σ^2_{err}). This comparison between the two models, (Eq. (1)) and (Eq. (3)), is also parallel to the situation of a two-level model, the typical fMRI group analysis where one takes the effect estimates from individual subjects without and with their sampling variances (Chen et al., 2012; Woolrich et al., 2004; Worsley et al., 2002).

Fitting the model of Eq. (3), briefly, is similar to a simpler case with a three-level model (instead of four-level) previously described by Konstantopoulos (2011) within the context of behavioral studies. In the present work, estimates of $\widehat{\beta}_{i(j(k))}$ and σ_i^2 were first generated by AFNI program 3dREMLfit (as described above in the Statistical Analysis subsection). Then they were provided as input to a customized R (https://www.R-project.org/) program that relies on R package *metaphor* (Viechtbauer, 2010) to compute voxel-wise estimates of $\sigma^2_{session}$, σ^2_{run}, and σ^2_{block} via an iterative algorithm that solved Eq. (3) via the restricted maximum likelihood scheme.

In addition, voxel-wise estimates of total variance were computed as the voxel-wise variance across all 500 beta estimates. Voxel-wise maps of σ^2_{err} were computed by averaging block-wise estimates of σ_i^2 generated by AFNI program *3dREMLfit* across all 500 blocks.

Network analysis

To evaluate potential differences in within-subject variance components across typical cognitive networks, we used previously published network maps from (Laird et al., 2011). This particular taxonomy was selected because clear behavioral correlates have been reported for each of the networks based on meta-analysis against task-based studies included in the *BrainMap* database (Fox et al., 2005). Four networks from the original taxonomy were excluded: two because they were originally identified as artifactual in the original study (networks 19 and 20) and two more because they do not fall completely within our imaging field of view. Table 1 shows detailed information regarding which networks were used, and the labeling scheme used for the remainder of this paper.

Network maps publicly available at the *BrainMap* website (http://www.brainmap.org/icns/) were brought from MNI space into each subject's specific space and converted to binary masks using a threshold of $Z>5$. Finally binary network maps were further restricted at the individual level to only contain grey matter voxels marked as significant in statistical maps of activation (*FDR* $q<0.05$) for the *Sustained Only Model* computed using all 100 runs in (Gonzalez-Castillo et al., 2012). For this purpose, grey matter ribbon masks were generated with the SPM segmentation tool using as input the high resolution anatomical scans of each subject. This last individual-level restriction was implemented to ensure that variance decomposition analyses were conducted only over voxels where a sustained response to the task was present. For completion, supplementary figures with maps containing all significantly active voxels (not only those within the grey matter ribbon) are also provided.

Temporal signal-to-noise ratio

Maps of voxel-wise temporal signal-to-noise ratio (TSNR) were computed for each run independently after the alignment step. Prior to

Table 1
Summary of networks.

Original network ID (Laird et al., 2011)	New network ID	Description
1	EI4	Emotion/Interoception Network #4
2	EI3	Emotion/Interoception Network #3
3	EI2	Emotion/Interoception Network #2
4	EI1	Emotion/Interoception Network #1
6	MV1	Motor/Visuospatial Network #1
7	MV2	Motor/Visuospatial Network #2
8	MV3	Motor/Visuospatial Network #3
9	MV4	Motor/Visuospatial Network #4
10	VS1	Visual Network #1
11	VS2	Visual Network #2
12	VS3	Visual Network #3
13	DMN	Default Mode Network
15	FPR	Right Fronto-Parietal Network
16	AUD	Auditory Network
17	SPP	Speech Production Network
18	FPL	Left Fronto-Parietal Network

computation of TSNR maps, task effects were regressed out, to avoid bias due to activity-induced fluctuations in TSNR values.

Per run voxel-wise TSNR maps were then averaged across all runs for each participant. These individual average TSNR maps were then used to compute representative TSNR values for each of the sixteen networks described above, separately for each participant. The TSNR value for each network is the average across all voxels part of that network.

Results

Fig. 1.A shows the spatial distribution of total within-subject variance for a representative subject (see Supplementary Fig. 1 and 2 for equivalent maps in the other two participants). Total variance is highest at the edge of the brain, the ventricles, inferior frontal regions commonly affected by dropout and B_o distortions, and occipital cortex (primary input for the task). Fig. 1B–E shows the spatial distribution for the variance decomposition analysis in the same representative subject. Measurement/modeling errors dominate across the brain and account for the majority of the within-subject variance (Fig. 1E). As for the other variance components, they show distinct spatial distributions. Both $\sigma^2_{session}$ and σ^2_{run} are highest at the edges of the brain. In addition, regions with high $\sigma^2_{session}$ and σ^2_{run} are present in occipital cortex. Finally, σ^2_{block} seems to be lowest across the majority of the brain, yet some clusters of high σ^2_{block} can be found in components of the default mode network.

To better understand the contribution of different variance components to somatosensory and cognitive networks, we constructed pie plots with their relative contributions to each network (Fig. 2). Somatosensory networks (i.e., MV1-4, VS1-3 and AUD) are in the top row of each subject's panel, while higher-order cognitive networks (EI1-4, DMN, FPR, FPL and SPP) are depicted below. Although the exact distribution of within-subject variance components across participants differs, several general patterns were observed. First, σ^2_{err} (white wedges) is the greatest contributor to within-subject variance in all networks. Second, σ^2_{run} (red wedges) is the second largest contributor to within-subject variance in all networks except VS2-3 for all subjects (only exception being VS3 in subject 2 where σ^2_{run} and $\sigma^2_{session}$ contribute similarly). Third, $\sigma^2_{session}$ (green wedges) is the second largest contributor only for early visual networks, which constitute the primary target of the experimental task in this dataset. Forth, σ^2_{block} (blue wedges) is a higher variance contributor in higher-order cognitive networks relative to somatosensory networks. This is particularly clear for subjects 1 and 3. Yet, for subject 2 the only three networks (EI1,3-4) where σ2block exceeds 1% are also higher-order cognitive networks. To evaluate the homogeneity of these profiles across different regions of a network, we decided to also compute median values for each variance subcomponent on a ROI-by-ROI basis. Supplementary Fig. 3 shows the results of this analysis. Despite some punctual differences in the median value of specific variance components across some intra-

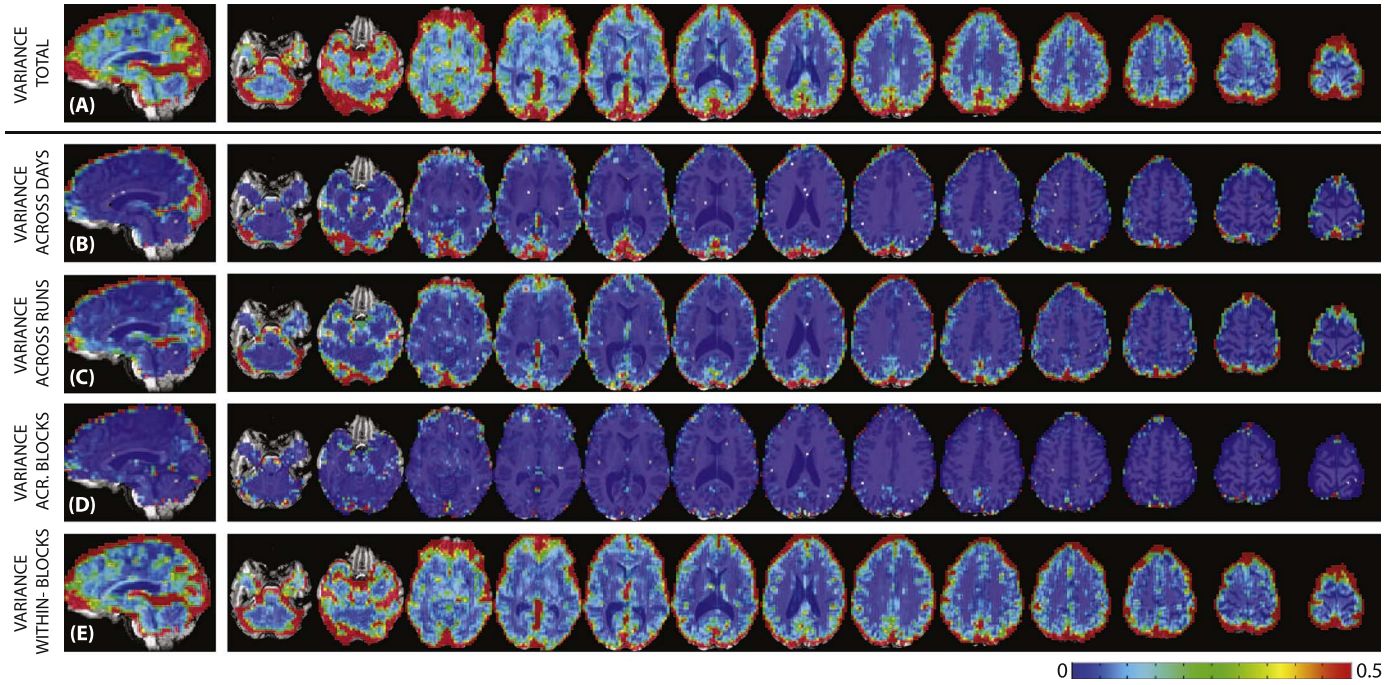

Fig. 1. Spatial Maps for Variance for a representative subject. (A) Total variance across the 500 estimates of effect size. (B) Variance across days. (C) Variance across runs within days. (D) Variance across blocks within runs and days. (E) Residual measurement/modeling error variance.

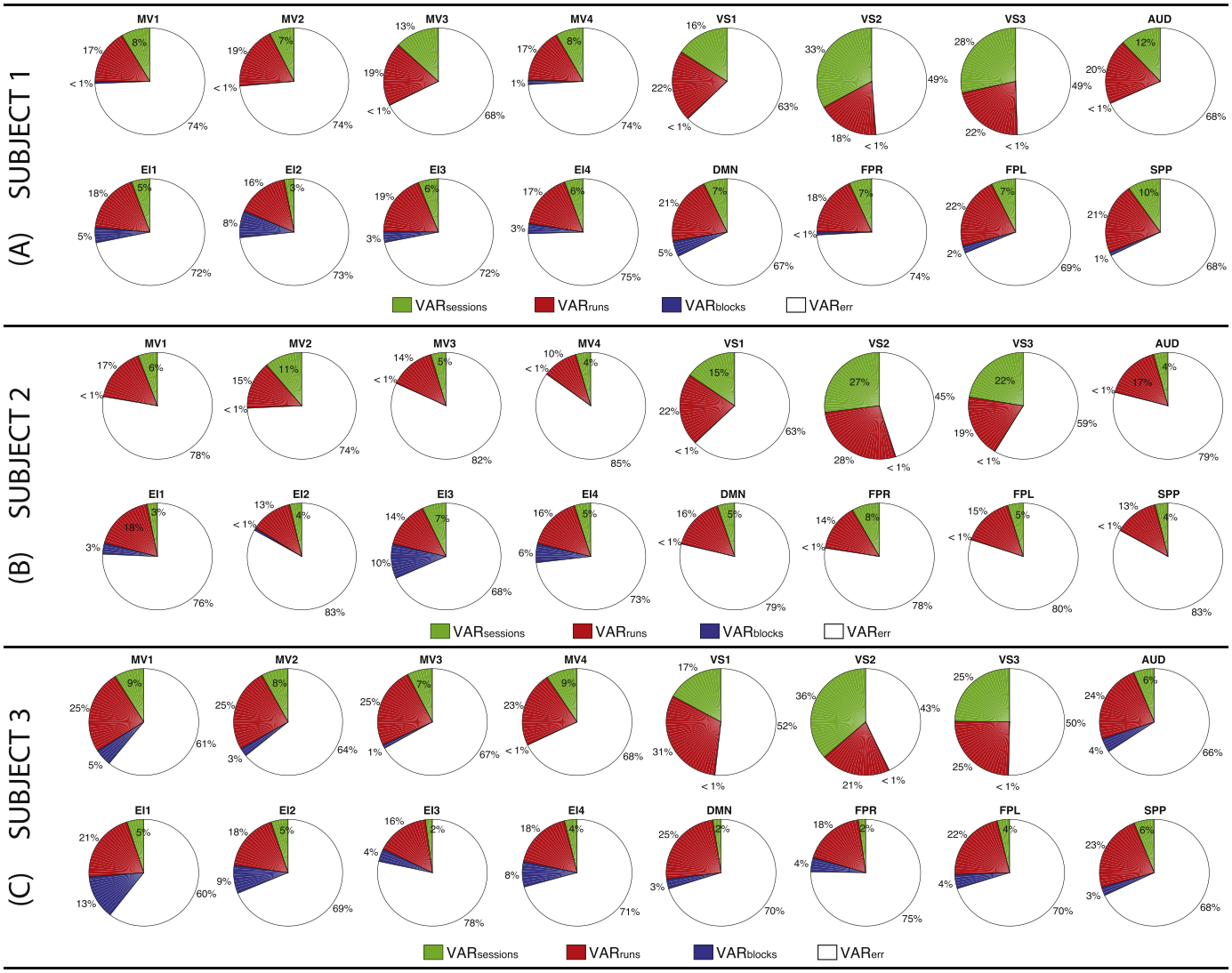

Fig. 2. Pie charts of median per-network contributions of the different variance components to total within-subject variance. (A) Subject 1. (B) Subject 2. (C) Subject 3. In all panels, somatosensory networks are depicted on the top row, and higher-order cognitive networks on the bottom row. When a component contributes less than 1%, a wedge for that component is absent. Color codes: $\sigma^2_{session}$ in green; σ^2_{run} in red, σ^2_{block} in blue, and σ^2_{err} in white.

network ROIs (e.g., DMN for all subjects, EI1 for subject 2), an overall agreement in the profile of variance decomposition across ROIs part of the same network could be observed.

Next, we focus our attention on $\sigma^2_{session}$, σ^2_{run}, and σ^2_{block} as their separate estimation constitutes the main novelty of this study. First, we explored the relationship between effect size and each of these three variance components on a network-by-network basis. Fig. 3 shows scatter plots of absolute values of median network-wise effect size against network-wise median estimates for the three variance components of interest: (A, B) variance-across sessions, (C, D) variance across runs within a session, and (E, F) variance across blocks within runs within sessions. Top panels (A, C, E) show all 16 networks and 3 subjects. Bottom panels (B, D, F) show the same information excluding the visual networks (VS1-3) to help better visualize the relationships for the other networks. In all plots, data points for subject 1 are represented as circles, for subject 2 as diamonds, and for subject 3 as squares (for a depiction of the same network-level variance decomposition results on a subject-by-subject basis please see Supplementary Fig. 4). The color of these symbols indicates the network. Warm colors are used to indicate higher-order cognitive networks—namely orange for EI1-4 and red for DMN, SPP, FPL, and FPR—while different shades of green is used to indicate somatosensory networks—dark green for VS1-3, green for MV1-4 and olive for AUD. Visual networks are characterized by higher average effect size than the rest of the networks. This higher effect size comes accompanied by higher across-day (Fig. 3A) and across-run (Fig. 3C), but not across-block (Fig. 3E) variance, compared to all other networks. Once visual networks are excluded, no significant linear relationship between effect size and any of the variance components was found (Fig. 3B, D and F; $p_{Bonf} > 0.05$ for all three attempted linear fits).

Next, we generated maps (Fig. 4; Supplementary Fig. 5) with voxels colored according to which is the second largest variance contributor (i.e., other than σ^2_{err}). In all subjects, σ^2_{run} (red) dominates across the majority of significantly active grey matter, with the exception of occipital cortex, where $\sigma^2_{session}$ (green) dominates. This confirms the network-based results of Fig. 2 at the voxel-wise level.

We also conducted a series of pair-wise variance component comparisons using three different variance ratios: $\sigma^2_{session}/\sigma^2_{run}$ (Fig. 5); $\sigma^2_{run}/\sigma^2_{block}$ (Fig. 6); and $\sigma^2_{session}/\sigma^2_{block}$ (Fig. 7). Figs. 5–7 contain voxel-wise maps of the above-mentioned ratios, as well as per-network percentages of voxels where the ratio is greater than one (i.e., the variance on the numerator is the largest) or less then one (i.e., the variance in the denominator is the largest).

When comparing $\sigma^2_{session}$ to σ^2_{run} via their ratio, we observe once more how $\sigma^2_{session}$ only dominates over σ^2_{run} in early visual networks/

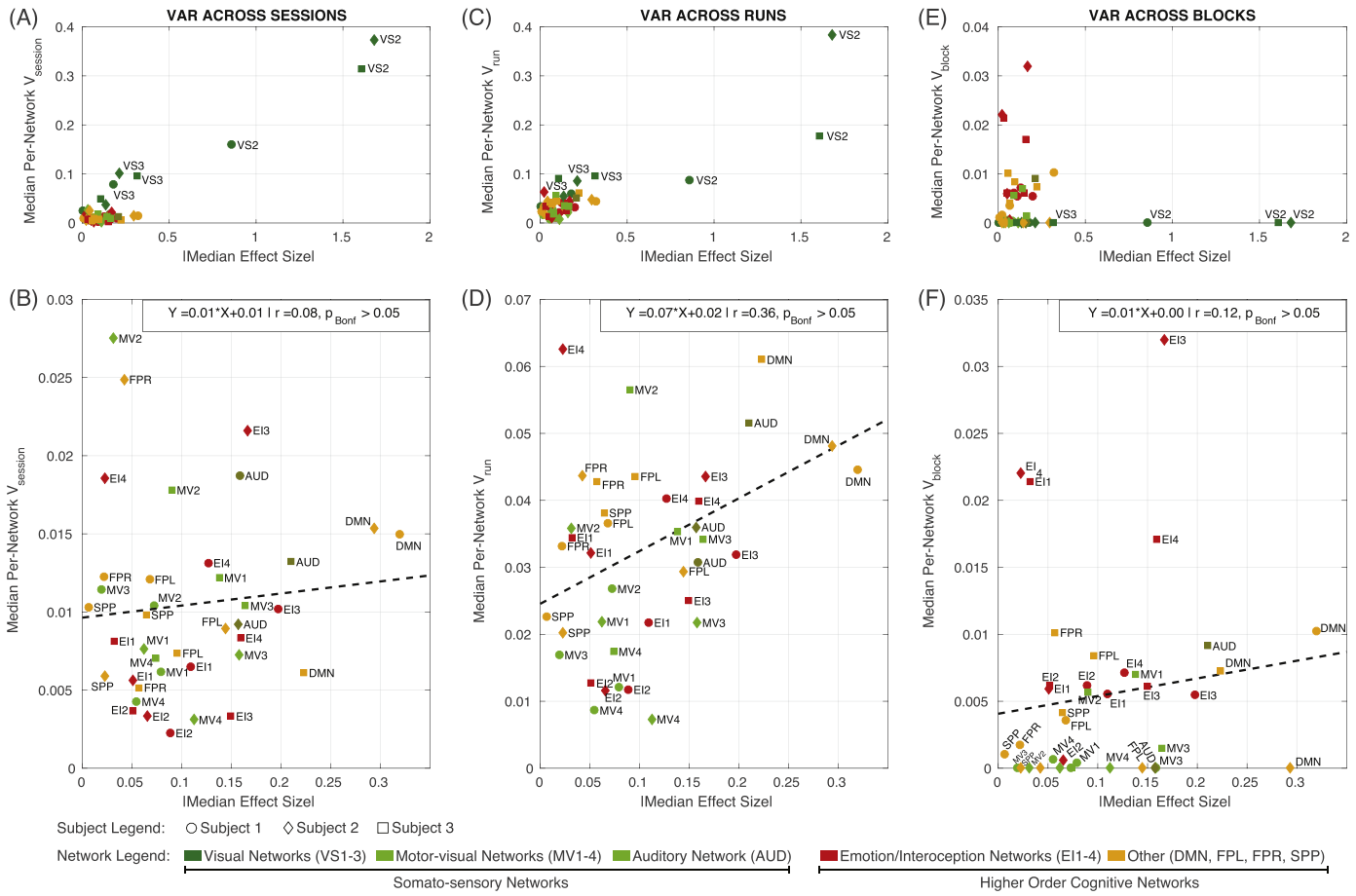

Fig. 3. Scatter plots of per-network median variance components against per-network absolute value of median effect size. (A, B) Across-session variance vs. effect size. (C, D) Across-run variance vs. effect size. (E, F) Across-block variance vs. effect size. Top panels (A, C, E) include all networks and subjects. Bottom panels (B, D, F) show the same information after excluding visual networks (VS1-3). In all panels, data points corresponding to subject 1 are depicted as circles, subject 2 as diamonds and subject 3 as squares. The color of these symbols indicates the type of network. Somato-sensory networks are depicted with different shades of green: dark green for visual networks (VS1-3), light green for motor-visual networks (MV1-4), and olive for auditory network (AUD). Higher-order cognitive networks are depicted with warm colors: red for emotion/interoception networks (EI1-4); and orange for the remaining higher-order networks (DMN, FPR, FPL, and SPP). Linear fitting attempts, none of which rendered significant at $p_{Bonf} < 0.05$, are shown as dashed black lines.

occipital cortex (Fig. 5A). The same is true in terms of within-networks voxel counts. Only for VS2 and VS3 the number of voxels with $\sigma^2_{session}/\sigma^2_{run} > 1$ account for more than 50% of voxels in the network (red dashed rectangle). Results for the other two subjects can be seen in Supplementary Fig. 6.

Similarly, in Fig. 6 (see Supplementary Fig. 8 for the other subjects) we can observe how, when σ^2_{run} and σ^2_{block} are compared directly to each other, σ^2_{run} dominates over σ^2_{block} in the majority of the brain. This is confirmed in the network-wise analysis for the $\sigma^2_{run}/\sigma^2_{block}$ ratio, which shows how, for all networks, voxels with a ratio greater than one account for the majority of within-network voxels.

Perhaps, the most interesting pair-wise comparison is that of $\sigma^2_{session}$ versus σ^2_{block} (Fig. 7; Supplementary Fig. 10). While $\sigma^2_{session}$ dominates over σ^2_{block} across most brain regions, in all subjects, we can observe how σ^2_{block} exceeds $\sigma^2_{session}$ in several subcortical regions, as well as nodes of the default mode network. For the particular instance

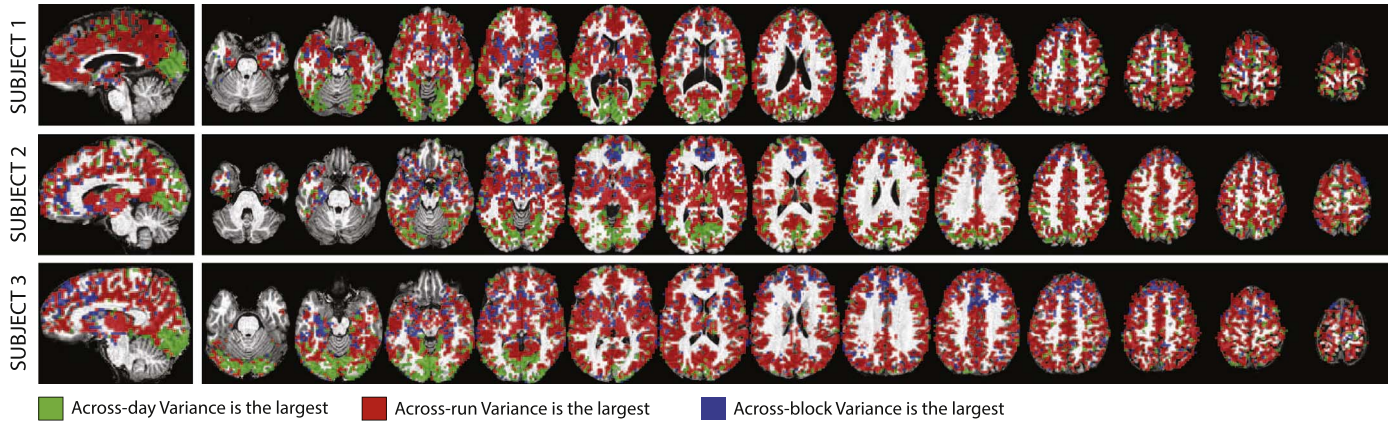

Fig. 4. Maps of largest variance component (excluding σ^2_{err}) for the three subjects under evaluation. Green signal voxels where $\sigma^2_{session}$ was the largest. Red signal voxels where σ^2_{run} was the largest. Finally, blue signal voxels were σ^2_{block} was the largest. Maps are restricted to significantly grey matter active voxels.

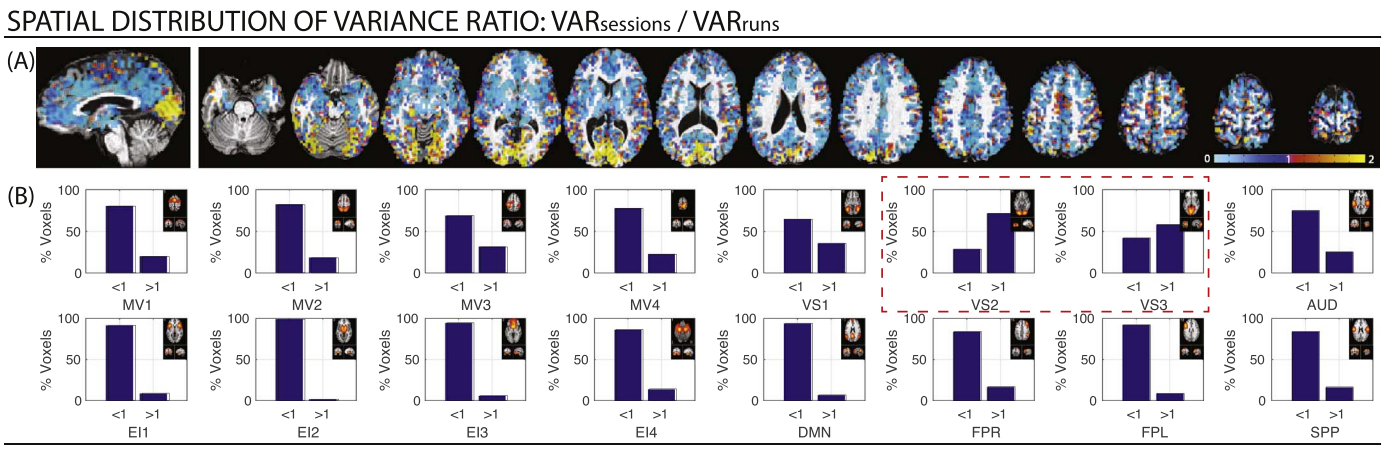

Fig. 5. (A) Maps of the voxel-wise $\sigma^2_{session}/\ \sigma^2_{run}$ ratio for a representative subject (Subject 1). Cold colors signal voxel where σ2session < σ2run, while hot colors signal voxels where $\sigma^2_{session} > \sigma^2_{run}$. (B) Percentage of within-network voxels where each of the variance dominates (i.e., ratio is above or below one) for the same subject. A small depiction of the original network definition in Laird et al. (2011) accompanies each bar graph for reference. Results for the other two subjects in Supplementary Fig. 6. In addition, the results of the same analyses when all significantly active voxels are taken into account are reported in Supplementary Fig. 7.

of subject 3, σ^2_{block} exceeds $\sigma^2_{session}$ also in several frontal locations. The network-wise analysis of the $\sigma^2_{session}/\sigma^2_{block}$ ratio revealed a reproducible pattern across subjects in which somatosensory networks (i.e., VS1-3, MV1-4 and AUD) contain predominantly voxels where $\sigma^2_{session}$ exceeds σ^2_{block}, while higher-order cognitive networks contain relatively larger proportions of voxels with $\sigma^2_{block}/\sigma^2_{session} > 1$ (red dashed rectangles). In some instances, such as network EI2 in subject 1, networks EI1, EI3 and EI4 for subject 2, and networks EI1-4, DMN, FPR and FPL for subject 3, voxels where σ^2_{block} exceeds $\sigma^2_{session}$ account for more than half of the network.

Fig. 8 shows individual averaged BOLD responses across all blocks and all significantly active grey matter voxels inside each network of interest. All networks, with the exception of FPL and SPP, show responses that follow, to different degrees, a sustained pattern of either positive or negative activity. Although networks show different, and in some cases prominent, deviations from the canonical expected response, it is not always the case that networks with the largest contribution of σ^2_{block} (Fig. 2) are the ones that are a worse fit for the canonical response. For example, subject 1 DMN and EI2, subject 2 EI3, and subject 3 DMN and EI4—all of which are networks with prominent σ^2_{block} contributions—; follow the canonical model better than subject 1 VS1, subject 2 SPP and VS1, and subject 3 MV4—which have almost no contribution from σ^2_{block} (Fig. 2).

Finally, to evaluate the influence of TSNR on the results, we computed average TSNR values per network in all three subjects. Fig. 9 shows the results of these analyses as bar plots. For each subject, networks are sorted by TSNR in descending order. For all subjects, higher-order cognitive networks (white bars) appear interleaved with somatosensory networks (dark-grey bars), suggesting there is not a clear relationship between TSNR and dominance of σ^2_{block} over $\sigma^2_{session}$.

Discussion

Variance of fMRI activity estimates is commonly decomposed into three random terms: measurement/modeling error, within-subject variance, and between-subject variance. Additional terms, such as between-site, may be added in studies that combine data across imaging centers (Sutton et al., 2008; Yendiki et al., 2010). Yet, a more accurate model is one that further subdivides within-subject variance into its three primary contributors: across-blocks (σ^2_{block}), across-runs (σ^2_{run}) and across-sessions ($\sigma^2_{session}$). Such finer model is difficult to estimate in practice because studies lack sufficient repeated within-subject measures under stable conditions (i.e., same task). One exception is the task-based dataset studied here. The large number of available intra-subject trials permitted us to segregate contributions due to measurement/modeling errors (σ^2_{err}) from those due to sessions, runs and blocks; and discover how these last three components (i.e., $\sigma^2_{session}$, σ^2_{run}, σ^2_{block}) have distinct spatially inhomogeneous distributions. And more specifically,

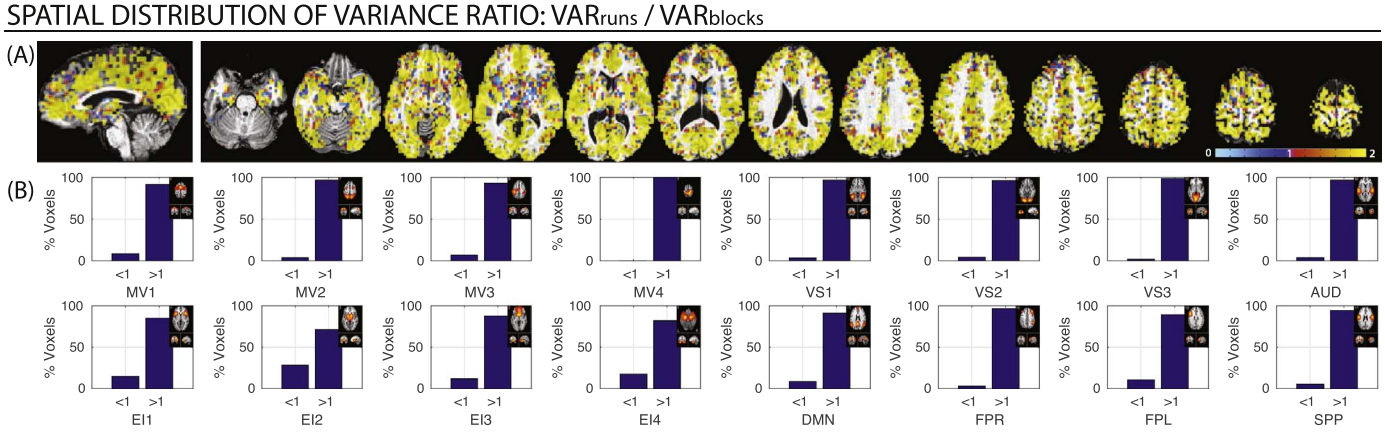

Fig. 6. (A) Maps of the voxel-wise σ2run/σ2block ratio for a representative subject (Subject 1). Cold colors signal voxel where σ2run < σ2block, while hot colors signal voxels where σ2run > σ2block. (B) Percentage of within-network voxels where each of the variance dominates (i.e., ratio is above or below one) for the same subject. A small depiction of the original network definition in Laird et al. (2011) accompanies each bar graph for reference. Results for the other two subjects in Supplementary Fig. 8. In addition, the results of the same analyses when all significantly active voxels are taken into account are reported in Supplementary Fig. 9.

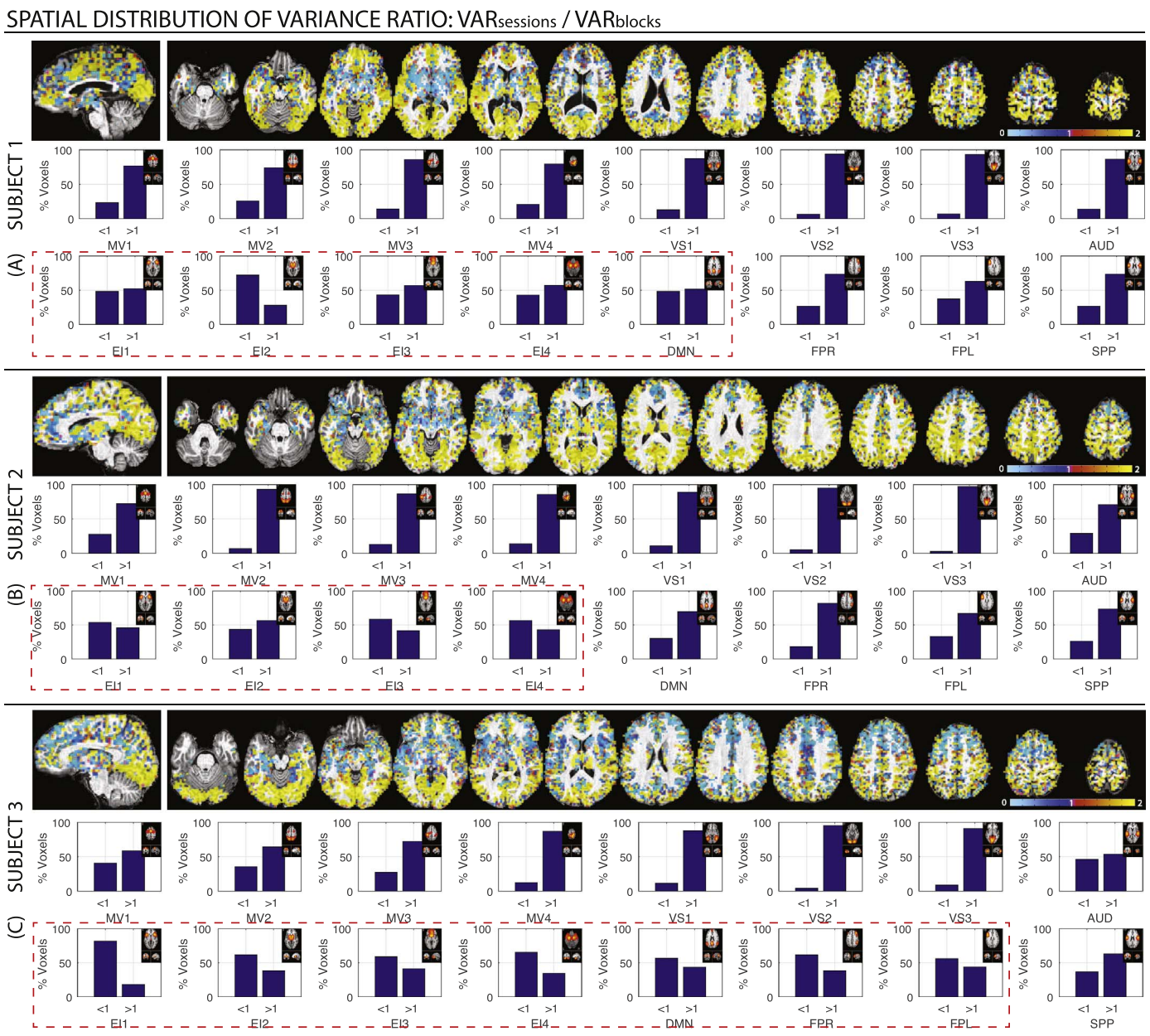

Fig. 7. Maps of voxel-wise $\sigma^2_{session}/\sigma^2_{block}$ ratios for all three subject. (A) Subject 1. (B) Subject 2. (C) Subject 3. For all subjects, on the top we show spatial maps of the voxel-wise ratio of interest. Cold colors signal voxel where $\sigma^2_{block} > \sigma^2_{session}$, while hot colors signal voxels where $\sigma^2_{session} > \sigma^2_{block}$. Below the maps, there are 16 bar plots per subject showing the percentage of voxels for which the ratio is greater and lower than one. Each bar plot corresponds to a different network.

how they contribute differently to the within-subject variance of somatosensory and higher-order cognitive networks.

Measurement/modeling error variance dominates across the brain

For all subjects, measurement/modeling error (σ^2_{err}) was the largest contributor to within-subject variance across the brain (Fig. 2). This was the case even after separating the effects of sessions, runs and blocks. Several prior studies have reported σ^2_{err} to be the largest variance contributor in fMRI (Friedman et al., 2006; Suckling et al., 2008) yet these previous accounts pooled variance across blocks and runs as part of the residual variance. Our results confirm that even when these contributions are properly segregated, σ^2_{err} remains the greatest source of within-subject variance across repeated measures.

Measurement error estimates include, in addition to random error, unexplained variance due to inaccurate modeling of expected responses. Hemodynamic responses are known to vary regionally within subject (Handwerker et al., 2004), yet few studies account for this variability. Moreover, they can have different relationships to task timing across the cortex (Gonzalez-Castillo et al., 2015, 2012; Uludağ, 2008), yet those are also commonly ignored. Given the prominent contribution of σ^2_{err} to within-subject variance everywhere in the brain, it follows that most substantial reductions in within-subject variance may result from additional efforts to account for inter-regional hemodynamic variability, as well as modeling of additional task components (e.g., transients at blocks onset and offsets). More generally, this result ultimately underscores the limitations of mass univariate General Linear Model (GLM) analyses for single-subject fMRI, which not only rely on spatial homogeneity of hemodynamic responses, but also on additional strong assumptions of linearity, including that of pure-insertion of cognitive processes when defining contrasts of interest (please see (Friston et al., 1996; Sartori and Umiltà, 2000) for a discussion on this particular topic). It may be that substantial reductions in σ^2_{err} at the single-subject level may only be

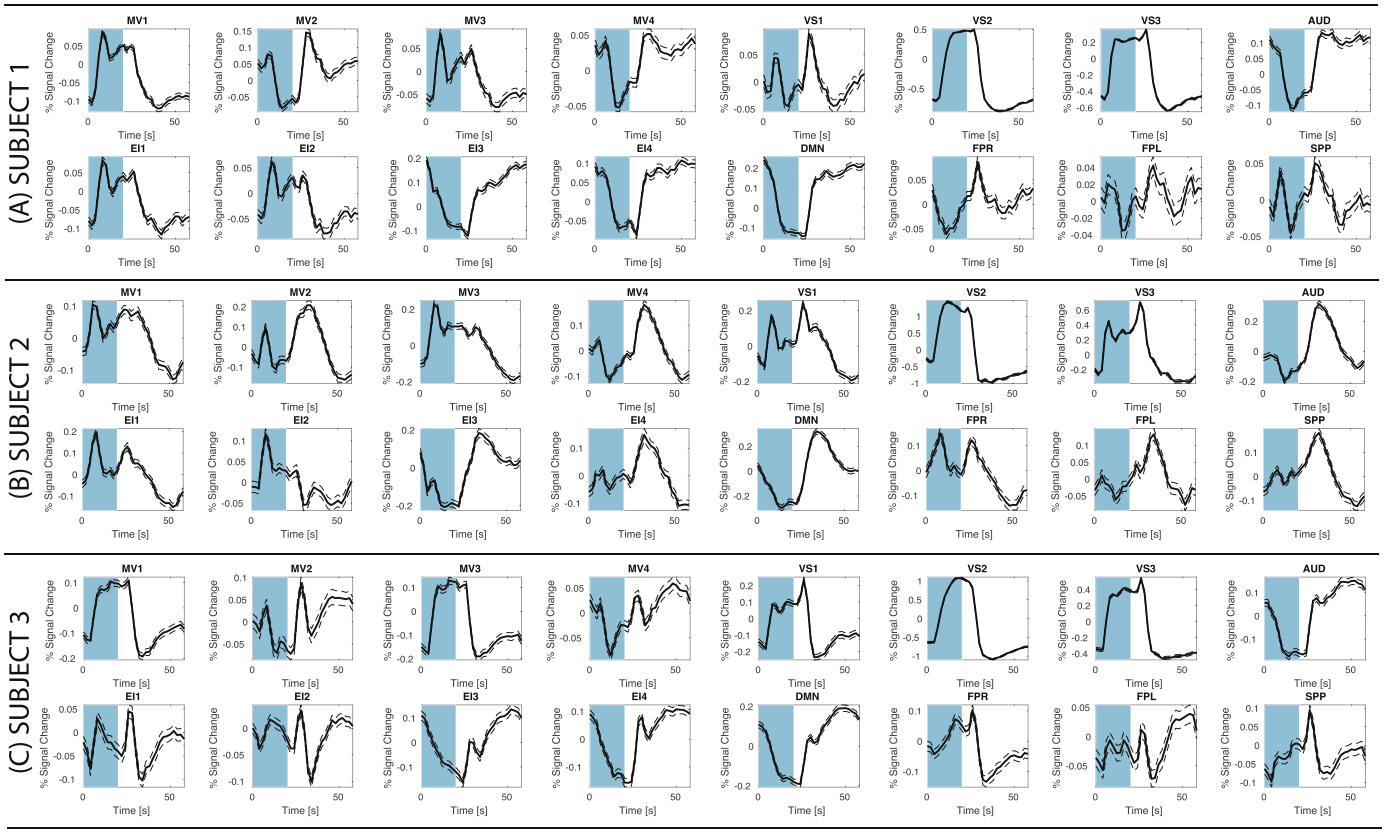

Fig. 8. Average block-wise responses across significantly active voxels in the sixteen networks of interest for all three subjects. (A) Subject 1. (B) Subject 2. (C) Subject 3. Each experimental block lasted 60 s, of which the initial 20 s (marked in cyan) corresponds to the active part of the block. In all plots, the average response is depicted with a continuous black line, while the standard error across these averages is shown as dotted traces.

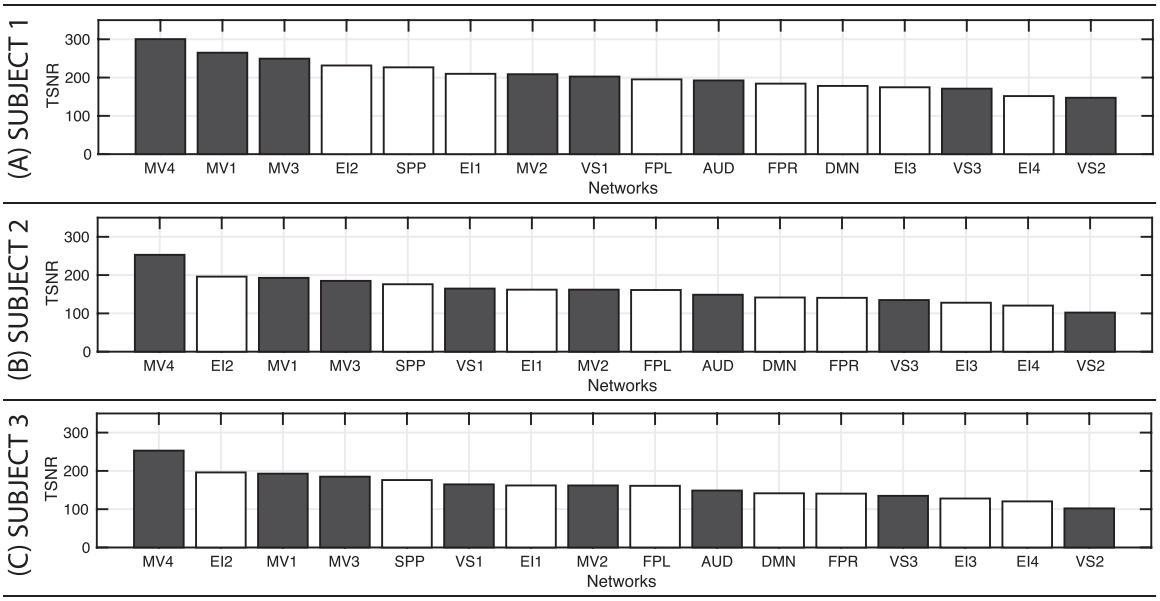

Fig. 9. Average Temporal Signal-to-Noise Ratio (TSNR) for all networks in each of the subjects. (A) Subject 1. (B) Subject 2. (C) Subject 3. For each subject networks are sorted by TSNR in descending order. TSNR for somatosensory networks are depicted as white bars. TSNR for higher-order cognitive networks as dark grey.

obtained via alternative multivariate data-driven analytical methods, such as Independent Component Analysis (Calhoun et al., 2001) or Self Organizing Maps (Katwal et al., 2013) that rely on a less stringent set of underlying assumptions about the data. Any such efforts may be vital to the success of longitudinal single-subject examinations with fMRI.

Most fMRI group analyses are conducted taking only into account individual effect size estimates, but no σ^2_{err} estimates, despite the availability of models and software(Chen et al., 2012)that can compute group-level statistics using both pieces of information. The prominence and spatial heterogeneity of σ^2_{err} as a contributor to within-subject variance reported here suggests that wider adoption of these advanced group level analytical methods may substantially improve group study results, as they can account for the inter-subject and inter-regional variability described here.

Across-session and across-runs contribute most prominently to visual networks

Across-session and across-run variances contributed approximately half of within-subject variance to the three visual networks (VS1: 41 ± 6%, VS2: 54 ± 3%, VS3: 47 ± 5%), while their joint average contribution across all other networks was approximately one forth (25 ± 4%). Moreover, across-session variance appeared to dominate over across-run variance in the majority of occipital cortex (Fig. 4; Supplementary Fig. 5). Different factors may have caused the elevated contribution of these two "longer-term" variance components.

First, visual networks had the strongest average response of all networks, which is expected for the task under examination. When response strength was plotted against different variance components, a clear relationship between response strength and across-session and across-run variance was observed (Fig. 3A and C) for visual networks, but not for across-block (Fig. 3E). This suggests that, to a given extent, larger values of across-session and across-run variance in visual networks are the result of larger responses in these regions.

Second, a potentially lower σ^2_{err} in absolute terms for the VS1-3 networks, relative to the rest of the brain, could render the relative contributions of any remaining sources of variance (e.g., $\sigma^2_{session}$) to appear disproportionally larger in these networks. Examination of absolute σ^2_{err} values (not shown) did not support this possibility. Moreover, hemodynamic responses (Fig. 8) and TSNR results (Fig. 9) for these networks also neglect it. Not all visual networks, and most particularly network VS1 in subjects 1 and 2, are either among the top TSNR networks or present hemodynamic responses that fit canonical standard sustained responses better than else in the brain; both of which could lead to a lower σ^2_{err} for these regions.

Third, it is possible that estimated responses in visual cortex do indeed present lower stability across repeated measures, especially across-sessions, relative to other regions. Factors previously shown to modulate occipital cortex responses to visual stimuli include: caffeine (Liu et al., 2004), attention(Jäncke et al., 1999; Specht et al., 2003), luminance (Liang et al., 2013), unstable fixation (Merriam et al., 2013), and even competing auditory stimulation, such as scanner noise (Zhang et al., 2005). None of these factors were appropriately controlled during the experiments (e.g., screen/mirror positioning may have varied across sessions resulting in differences in luminance (Strasburger et al., 2002), and therefore they should be considered likely contributors to across-session, and in some instances, also across-run variance in visual regions. Yet, many of these factors are known to also modulate activity outside visual regions. It is therefore not easy to discern whether observed elevated contributions of across-session and across-run variance in visual networks are the result of different contributing factors affecting different regions (e.g., factor A adds variance across-sessions in VS1 but not the DMN), inter-regional differences in contribution levels of the same factor (e.g., factor A affects activity levels to a larger extent in VS1 than in DMN), or a combination of both.

Finally, the elevated within-subject across-session variance in visual regions reported here for a task-based dataset is in agreement with the results from two separate *high-N* ($N_{sessions}$=158 and $N_{sessions}$=84 respectively) within-subject longitudinal evaluations of connectivity using resting-state scans (Choe et al., 2015; Poldrack et al., 2015). In both of these studies, visual networks were reported to be among those with the greatest degree of within-subject variability across sessions. This suggests that visual regions are characterized by high across-session within-subject variability independently of whether or not these regions are being driven by external task demands.

Across-blocks variance contributed more prominently to higher-order cognitive networks

Across-block variance was the smallest contributor of variance to all networks in all subjects (Fig. 2). This is not surprising given the closer temporal proximity of items contributing to this variance (seconds to a few minutes apart) relative to the other two "longer-term" variance components (i.e., $\sigma^2_{session}$ and σ^2_{run}). Moreover, given that physiological noise corrections are performed on a run-by-run basis, within-run blocks can be expected to have residual levels of physiological noise that are more similar than different runs do (e.g., due to differences in quality of physiological recordings across runs). Similarly, there is a higher probability of substantially larger head repositioning between than within runs (average within-run maximum volume-to-volume displacement=0.42 ± 0.23; average within-day across-run displacement=1.15 ± 0.92); making differences in geometric distortions a potential lower contributor to σ^2_{block} as well. For all subjects, spatial maps for the different variance components (Fig. 1, Supplementary Figs. 1 and 2) confirm these hypotheticals as they show how σ^2_{block} is smaller than σ^2_{block} or σ^2_{block} at the edges of the brain, ventricles and near prominent vascular structures.

Voxel-wise maps of $\sigma^2_{session}/\sigma^2_{block}$ (Fig. 7) tentatively indicate a greater contribution of σ^2_{block} to regions embedded in higher-order cognitive networks, particularly in subcortical regions (all subjects) and components of the default mode network (subjects 1 and 3). Yet, the unsmoothed and noisy profile of these voxel-wise maps make ascertaining any clear inferences difficult. A sharper picture emerges when analyses are conducted at the network-level. Despite the low contribution of σ^2_{block} to within-subject variance for all networks, we were able to detect an interesting trend across all subjects, namely that across-block variance contributes more to total within-subject variance in higher-order cognitive networks (4.0 ± 1.8%) than in somatosensory networks (0.6 ± 0.9 %). Moreover, for a subset of those higher-order cognitive networks (more specifically those labeled EI1-4), voxels with $\sigma^2_{block} > \sigma^2_{session}$ accounted for approximately half of intra-network significantly active grey matter voxels (Fig. 7; Subject 1: 51.5 ± 13.9%, Subject 2: 53.2 ± 6.7%, Subject 3: 66.9 ± 10.3). In two subjects, this behavior also extended to the DMN. Laird et al. (2011) originally described the networks labeled here EI1-4 as being strongly related to a collective range of emotional, interoceptive and autonomic processes. In the same study, the network labeled as DMN was associated with theory of mind and social cognition tasks, when contrasted against the *BrainMap* database. Although all these cognitive processes are to a large extent tangential to our task (e.g., our task had no emotional or social content), significant responses, both positive and negative, were detected when sufficient CNR was available. It is possible that high across-block variability for these regions is a consequence of such a loose relationship between our task processing requirements and what are thought to be the main functional roles of these regions. Moreover, in our original study we stated that the detection of brain-wide activations in fMRI (when CNR is sufficiently high) poses a very difficult question: *"…if a task-driven BOLD response is triggered across the whole brain, how does one differentiate between BOLD responses from regions critical for handling the task, versus regions that are not?"* It is possible that detailed variance analysis such as the ones reported here may help answer this question if for example regions not essential to task performance were to be reliably and distinctly characterized by across-block variance that exceeds across-session variance. We hope future work can help test the validity of this speculative, yet potentially powerful, notion.

Factors contributing to natural within-subject variance

Potential sources of longitudinal within-subject variance in fMRI recordings include, but are not limited to: habituation effects (Hamid et al., 2015), strategy shifts/practice effects (Kelly and Garavan, 2005), fatigue, lapses of attention, caffeine (Koppelstaetter and Poeppel, 2010; Liu et al., 2004), nicotine (Warbrick et al., 2012, 2011), time-of-day (Gaggioni et al., 2014; Schmidt et al., 2015), aging (Cliff et al., 2013; Koch et al., 2010), residual levels of physiological noise(Birn, 2012),

distinct geometric distortions across sessions (Raemaekers et al., 2012), or progression of clinical conditions. As our understanding of natural within-subject variability in both neuronal and fMRI responses improves, additional factors may need to be added to this list.

Although it is difficult to conclusively evaluate the potential contribution of all these sources to our variance decomposition, several factors can be ruled unlikely given the experimental tasks and procedures. The dataset reported here was collected over a time span of approximately 3 months in healthy young individuals. Therefore, aging, cognitive decline and disease can be excluded with a high degree of confidence. Practice effects are also unlikely given the simplicity of the task and the consistently high performance revealed by concurrent behavioral metrics (above 95% accuracy; see Gonzalez-Castillo et al. (2012) for additional details.). Similarly, evaluation of average positive response levels in VS3 (Supplementary Fig. 11), which includes primary visual cortex, did not show any clear pattern of habituation across sessions (i.e., monotonous decrease in activation as days progresses), making this factor also an improbable contributor of variance. Regarding time-of-day effects, although all scans were not always conducted at the same time, 86% of scans happened in the afternoon between noon and 6 p.m., with the remaining happening at later hours of the day (never concluding after 10 p.m.). As such, time-of-day effects might be considered negligible. Finally, only one subject reported to be a smoker. Given that similar levels of variability were observed in all participants, levels of nicotine consumption can also be thought as an unlikely contributor to within-subject variance here.

Other factors such as fatigue, variable attention, caffeine, residual misalignment and physiological noise are more likely to be among the strongest contributors to observed variance here. Caffeine has been shown to significantly affect the shape and duration of hemodynamic responses in visual cortex using a stimulus of very similar characteristics to ours(Liu et al., 2004). Given that we did not control for caffeine consumption in the hours preceding each scanning session, it is possible that caffeine levels may have been a contributing factor here. Regarding residual physiological noise and misalignment, our data suggest that these have also contributed to the results, despite our best efforts at accounting for them during pre-processing. Spatial maps of within-subject variance (Fig. 1, Supplementary Figs. 1 and 2) show large contributions from $\sigma^2_{session}$ and σ^2_{run} both in the edges of the brain—signaling residual motion or misalignment—as well as in the ventricles and large vascular structures (e.g., Circle of Willis), which suggests contributions from residual physiological noise. Finally, the experimenters visually confirmed the presence of clear positively sustained activation in primary visual cortex for all 1500 blocks. Such visual confirmation, combined with the high accuracy reported for the letter/number discrimination task, suggests that subjects attended to the stimuli and were compliant with the task in all instances. Yet, it does not preclude fatigue, shifts in motivation and short attention lapses to have contributed variance to the data. This is particularly true considering the highly repetitive and monotonous nature of our task.

A better characterization of contributing variance could be obtained if per-session phenotypic information, such as in (Poldrack et al., 2015) were available. Unfortunately that is not the case for the dataset studied here. Several institutions have started, or are currently in the process, of collecting large publicly available fMRI dataset, yet the focus is mainly on resting-state and large samples of subjects (Essen et al., 2012; Yan et al., 2013). While these datasets are an invaluable asset in our quest for uncovering fundamental principles of the structural and functional organization of the human brain, they are limited when it comes to obtaining a better understanding of natural—i.e., to be expected in the absence of any clinical development—within-subject variability of fMRI responses to task and its contributing sources. We believe that the parallel acquisition and publication of highly-sampled, multi-task, single-subject fMRI datasets annotated with phenotype-wide session specific information may be an equally valuable contribution to our understanding of the brain. Such datasets will provide new insights into the brain's natural variability in response to external stimulation and cognitive challenges. Moreover, in a time when many fMRI groups are turning their attention from studying commonalities in activity and connectivity patterns across pseudo-homogenous populations (e.g., healthy adults, autism, etc.) to finding optimal ways to capture those aspects of fMRI that are unique to each subject (Finn et al., 2015; Laumann et al., 2015; Poldrack et al., 2015), getting such a detailed understanding of within-subject natural variability is a fundamental step. Finally, such a dataset can also help inform the future development of fMRI clinical protocols. Although, many scientists and clinicians alike foresee resting-state as the primary paradigm for clinical fMRI (Khanna et al., 2015; Shimony et al., 2009), task-based fMRI is also clinically relevant, as clearly evidenced by its inclusion in many existing pre-surgical protocols (Hirsch et al., 2000; Stippich et al., 2007). Low test-retest reliability is often cited as a reason why fMRI has not been widely adopted in clinical practice (Stevens et al., 2016). Understanding and modeling naturally occurring, clinically irrelevant within- and between-subject variance is key to improving its reproducibility, and with it, its suitability for the clinic.

Limitations of the study

In our original analyses of this dataset we focused on the commonalities of responses across all blocks and discovered that small, yet meaningful, responses could be found in the majority of the brain. Here, we focused on the differences and attempted a within-subject variance decomposition analysis. Yet, some of the original limitations remain. First, despite having a larger-than-usual number of samples per subject, we have a very limited set of subjects. Although our conclusions are based only on those patterns of variance that were consistent across all subjects, the sample remains too small to make any generalizations or perform adequate statistical analysis to support more specific conclusions. Second, all subjects performed the same experimental task, precluding any evaluation of generalization of observations to other tasks (Plichta et al., 2012). Future studies with tasks targeting other sensory and cognitive systems will help elucidate if the spatial patterns of variance reported here are generalizable across tasks—and therefore represent fundamental principles of how components of within-subject variance appear in the brain—or if they are task-dependent (e.g., should higher-order cognitive networks be always expected to have higher across-block natural variability in their responses given their putative roles, or can such variability be modulated by task demands?). Third, analyses reported here focused solely on response estimates obtained using a single sustained canonical response model; despite evidence that responses with different temporal profiles (e.g., onset/offset only responses) are present (Gonzalez-Castillo et al. 2012). The use of more versatile models that allow for additional response types will affect variance components estimates (e.g., measurement error should decrease), and in turn may affect the relative contributions reported here. We focused here on sustained responses because these are the ones commonly reported in the literature for block design experiments. Additional analyses should evaluate the effect of modeling decisions on the within-subject variance decompositions reported here. It is also worth mentioning that the 40s off periods used in the present study are not the most common practice in block-designs, and that it is possible that offset durations may modulate observed variability patterns. Fourth, our analyses focus solely on the decomposition of variance for effect size estimates. Additional analyses should evaluate if variance decomposes equally for other activity summary metrics such as activation extent, activation overlap, etc. Fifth, all sessions were acquired within a period of three months. It is possible that if data were acquired over longer periods the contribution of across-sessions variance may increase. Longer longitudinal evaluations will be needed to answer this limitation.

Previous studies that have evaluated the temporal evolution of

within-subject variance for connectivity estimates at different temporal scales have found meaningful, spatially inhomogeneous, non-artifactual dynamic changes (Choe et al., 2015; Gonzalez-Castillo et al., 2014, 2015) that help inform the analysis and interpretation of longitudinal single-subject resting fMRI studies. Similarly, understanding the relative contributions of blocks, runs and sessions to within-subject variance can guide how to best combine and interpret longitudinal single-subject task-based results. It can also help optimize protocols and data acquisition for longitudinal studies, both for clinical and research purposes. We believe that despite the limitations cited above, the exploratory analyses reported here constitute a first step in that direction and will instigate working hypotheses for future more detailed evaluations of natural within-subject variability of fMRI responses.

Conclusions

Within-subject variance for effect size estimates of activity was decomposed in four nested components: across-sessions, across-runs within sessions, across-blocks within runs within sessions, and residual variance. Exploration of the contribution of these variance components to sixteen brain networks provided new insights on how individual subject variance is distributed spatially across the brain and temporally across these primary experimental units (i.e., blocks, runs, sessions and error). In particular, we showed that measurement error is the dominant source of within-subject variance across the brain even when variance across-blocks, runs and sessions are properly accounted for. Next, we showed that the second dominant source of variance for visual regions is across-sessions variance, while for the rest of the brain it was across-runs variance. Finally, we showed how across-block variance is a larger contributor of naturally occurring within-subject variance in high-order cognitive networks relative to that of somatosensory networks. These results suggest that efforts to minimize within-subject variability of activity estimates in single-subject examinations should focus primarily on reducing measurement error (e.g., use of more accurate response models that account for spatial and temporal heterogeneity of hemodynamic responses). In addition, the elevated contribution of across-block variance to higher-order cognitive networks suggests that these networks respond in a less reliable manner across blocks relative to primary somatosensory networks (at least within the context of the current task). As such, stable characterization of higher-order cognitive regions in individual subjects (e.g., for longitudinal and or clinical purposes) will require more samples than that of primary somatosensory regions.

Acknowledgements

This research was possible thanks to the support of the National Institute of Mental Health Intramural Research Program. We gratefully acknowledge the advice of Wolfgang Viechtbauer on variance decomposition and the usage of his R package *metafor*. Portions of this study used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). This study is part of NIH clinical protocol number NCT00001360, protocol ID 93-M-0170 and annual report ZIAMH002783-14. The research and writing of the paper were also supported by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHS, USA.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuroimage.2016.10.024.

References

Bianciardi, M., Fukunaga, M., Gelderen, P. van, 2009. Sources of functional magnetic resonance imaging signal fluctuations in the human brain at rest: a 7 T study. Magn. Reson. Imaging 27, 1019–1029.
Birn, R., Smith, M., Jones, T., Bandettini, P., 2008. The respiration response function: the temporal dynamics of fMRI signal fluctuations related to changes in respiration. Neuroimage 40, 644–654.
Birn, R.M., 2012. The role of physiological noise in resting-state functional connectivity. Neuroimage 62, 864–870.
Calhoun, V.D., Adali, T., Pearlson, G.D., Pekar, J.J., 2001. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. Hum. Brain Mapp. 13, 43–53.
Chen, G., Saad, Z.S., Nath, A.R., Beauchamp, M.S., Cox, R.W., 2012. FMRI group analysis combining effect estimates and their variances. Neuroimage 60, 747–765.
Choe, A., Jones, C., Joel, S., Muschelli, J., Belegu, V., Caffo, B., Lindquist, M., Zijl, P., van, Pekar, J., 2015. Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years. PLoS One 10, e0140134.
Cliff, M., Joyce, D., Lamar, M., Dannhauser, T., Tracy, D., 2013. Aging effects on functional auditory and visual processing using fMRI with variable sensory loading. Cortex 49, 1304–1313.
Cox, R., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.
Dayan, E., Cohen, L.G., 2011. Neuroplasticity subserving motor skill learning. Neuron.
Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., 2012. The human connectome project: a data acquisition perspective. Neuroimage 62, 2222–2231.
Finn, E., Shen, X., Scheinost, D., Rosenberg, M., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671.
Fox, P.T., Laird, A.R., Fox, S.P., Fox, M.P., Uecker, A.M., Crank, M., Koenig, S.F., Lancaster, J.L., 2005. BrainMap taxonomy of experimental design: description and evaluation. Hum. Brain Mapp. 25, 185–198.
Friedman, L., Glover, G., Consortium, F., 2006. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. NeuroImage 33, 471–481.
Friston, K., Price, C., Fletcher, P., Moore, C., 1996. The trouble with cognitive subtraction. Neuroimage 4, 97–104.
Gaggioni, G., Maquet, P., Schmidt, C., Dijk, D., 2014. Neuroimaging, cognition, light and circadian rhythms. Front. Syst. Neurosci. eCollect., 2014.
Glover, G., Li, T., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: retroicor. Magn. Reson. Med. 44, 162–167.
Gonzalez-Castillo, J., Talavage, T., 2011. Reproducibility of fMRI activations associated with auditory sentence comprehension. NeuroImage 54, 138–155.
Gonzalez-Castillo, J., Saad, Z., Handwerker, D., Inati, S., Brenowitz, N., Bandettini, P., 2012. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proc. Natl. Acad. Sci. USA 109, 5487–5492.
Gonzalez-Castillo, J., Handwerker, D.A., Robinson, M.E., Hoy, C.W., Buchanan, L.C., Saad, Z.S., Bandettini, P.A., 2014. The spatial structure of resting state connectivity stability on the scale of minutes. Front. Neurosci. 8, 138.
Gonzalez-Castillo, J., Hoy, C.W., Handwerker, D.A., Roopchansingh, V., Inati, S.J., Saad, Z.S., Cox, R.W., Bandettini, P.A., 2015. Task dependence, tissue specificity, and spatial distribution of widespread activations in large single-subject functional MRI datasets at 7 T. Cereb. Cortex 25, 4667–4677.
Gountouna, V., Job, D., McIntosh, A., Moorhead, T., 2010. Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49, 552–560.
Hamid, A., Speck, O., Hoffmann, M., 2015. Quantitative assessment of visual cortex function with fMRI at 7 Tesla-test-retest variability. Front. Human. Neurosci., 9.
Handwerker, D., Ollinger, J., D'Esposito, M., 2004. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. Neuroimage 21, 1639–1651.
Havel, P., Braun, B., Rau, S., Tonn, J., Fesl, G., 2006. Reproducibility of activation in four motor paradigms. J. Neurol. 253, 471–476.
Hirsch, J., Ruge, M.I., Kim, K.H., Correa, D.D., Victor, J.D., Relkin, N.R., Labar, D.R., Krol, G., Bilsky, M.H., Souweidane, M.M., DeAngelis, L.M., Gutin, P.H., 2000. An integrated functional magnetic resonance imaging procedure for preoperative mapping of cortical areas associated with tactile, motor, language, and visual functions. Neurosurgery 47, 711–721.
Jäncke, L., Mirzazade, S., Shah, N., 1999. Attention modulates the blood oxygen level dependent response in the primary visual cortex measured with functional magnetic resonance imaging. Naturwissenschaften 86, 79–81.
Jo, H., Saad, Z.S., Simmons, K.W., Milbury, L.A., Cox, R.W., 2010. Mapping sources of correlation in resting state FMRI, with artifact detection and removal. Neuroimage 52, 571–582.
Katwal, S.B., Gore, J.C., Marois, R., Rogers, B.P., 2013. Unsupervised spatiotemporal analysis of fMRI data using graph-based visualizations of self-organizing maps. IEEE Trans. Bio-Med. Eng. 60, 2472–2483.
Kelly, C.A., Garavan, H., 2005. Human functional neuroimaging of brain changes associated with practice. Cereb. Cortex 15, 1089–1102.
Khanna, N., Altmeyer, W., Zhuo, J., Steven, A., 2015. Functional neuroimaging: fundamental principles and clinical applications. Neuroradiol. J 28, 87–96.
Koch, W., Teipel, S., Mueller, S., Buerger, K., Bokde, A., 2010. Effects of aging on default mode network activity in resting state fMRI: does the method of analysis matter? Neuroimage 51, 280–287.

Konstantopoulos, S., 2011. Fixed effects and variance components estimation in three-level meta-analysis. Res. Synth. Methods 2, 61–76.
Koppelstaetter, F., Poeppel, T., 2010. Caffeine and cognition in functional magnetic resonance imaging. J. Alzheimers Dis. 20, S71–S84.
Laird, A., Fox, M., Eickhoff, S., Turner, J., Ray, K., McKay, R., Glahn, D., Beckmann, C., Smith, S., Fox, P., 2011. Behavioral interpretations of intrinsic connectivity networks. J. Cogn. Neurosci. 23, 4022–4037.
Laumann, T., Gordon, E., Adeyemo, B., Snyder, A., Joo, S., Chen, M.-Y., Gilmore, A., McDermott, K., Nelson, S., Dosenbach, N., Schlaggar, B., Mumford, J., Poldrack, R., Petersen, S., 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron 87, 657–670.
Liang, C.L., Ances, B.M., Perthen, J.E., Moradi, F., Liau, J., Buracas, G.T., Hopkins, S.R., Buxton, R.B., 2013. Luminance contrast of a visual stimulus modulates the BOLD response more than the cerebral blood flow response in the human brain. Neuroimage 64, 104–111.
Liu, T., Behzadi, Y., Restom, K., Uludag, K., Lu, K., 2004. Caffeine alters the temporal dynamics of the visual BOLD response. Neuroimage 23, 1402–1413.
Luck, S.J., 2014. An Introduction to the Event-related Potential Technique. MIT Press, Cambridge, MA.
McGonigle, D., 2012. Test–retest reliability in fMRI: or how I learned to stop worrying and love the variability. NeuroImage 62, 1116–1120.
McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R., Holmes, A.P., 2002. Variability in fMRI: an examination of intersession differences. Neuroimage 11, 708–734.
McKenna, B., Drummond, S., Eyler, L., 2014. Associations between circadian activity rhythms and functional brain abnormalities among euthymic bipolar patients: a preliminary study. J. Affect. Disord. 164, 101–106.
Merriam, E.P., Gardner, J.L., Movshon, A.J., Heeger, D.J., 2013. Modulation of visual responses by gaze direction in human visual cortex. J. Neurosci. 33, 9879–9889.
Plichta, M., Schwarz, A., Grimm, O., Morgen, K., Mier, D., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60, 1746–1758.
Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.-Y.Y., Gorgolewski, K.J., Luci, J., Joo, S.J., Boyd, R.L., Hunicke-Smith, S., Simpson, Z.B., Caven, T., Sochat, V., Shine, J.M., Gordon, E., Snyder, A.Z., Adeyemo, B., Petersen, S.E., Glahn, D.C., Reese Mckay, D., Curran, J.E., Göring, H.H.H., Carless, M.A., Blangero, J., Dougherty, R., Leemans, A., Handwerker, D.A., Frick, L., Marcotte, E.M., Mumford, J.A., 2015. Long-term neural and physiological phenotyping of a single human. Nat. Commun. 6, 8885.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154.
Raemaekers, M., Plessis, D.S., Ramsey, N., 2012. Test–retest variability underlying fMRI measurements. Neuroimage 60, 717–727.
Sartori, G., Umiltà, C., 2000. How to avoid the fallacies of cognitive subtraction in brain imaging. Brain Lang. 74, 191–212.
Schmidt, C., Collette, F., Reichert, C., Maire, M., 2015. Pushing the limits: chronotype and time of day modulate working memory-dependent cerebral activity. Front. Neurol. 6, 199.
Shimony, J.S., Zhang, D., Johnston, J.M., Fox, M.D., Roy, A., Leuthardt, E.C., 2009. Resting-state spontaneous fluctuations in brain activity: a new paradigm for presurgical planning using fMRI. Acad. Radiol. 16, 578–583.
Skoe, E., Kraus, N., 2010. Auditory brain stem response to complex sounds: a tutorial. Ear Hear. 31, 302–324.
Specht, K., Willmes, K., Shah, J.N., Jäncke, L., 2003. Assessment of reliability in functional imaging studies. J. Magn. Reson. Imaging 17, 463–471.
Stevens, M.T., Clarke, D.B., Stroink, G., Beyea, S.D., D'Arcy, R.C., 2016. Improving fMRI reliability in presurgical mapping for brain tumours. J. Neurol. Neurosurg. Psychiatry 87, 267–274.
Stippich, C., Rapps, N., Dreyhaupt, J., Durst, A., Kress, B., Nennig, E., Tronnier, V.M., Sartor, K., 2007. Localizing and lateralizing language in patients with brain tumors: feasibility of routine preoperative functional MR imaging in 81 consecutive patients. Radiology 243, 828–836.
Strasburger, H., Wüstenberg, T., Jäncke, L., 2002. Calibrated LCD/TFT stimulus presentation for visual psychophysics in fMRI. J. Neurosci. Methods 121, 103–110.
Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S., Graves, M., Chen, C., Spiegelhalter, D., Bullmore, E., 2008. Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum. Brain Mapp. 29, 1111–1122.
Sutton, B., Goh, J., Hebrank, A., Welsh, R., 2008. Investigation and validation of intersite fMRI studies using the same imaging hardware. J. Magn. Reson. Imaging 28, 21–28.
Uludağ, K., 2008. Transient and sustained BOLD responses to sustained visual stimulation. Magn. Reson. Imaging 26, 863–869.
Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. 36, 1–48.
Vuilleumier, P., Driver, J., 2007. Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. Philos. Trans. R. Soc. Lond. B Biol. Sci. 362, 837–855.
Warbrick, T., Mobascher, A., Brinkmeyer, J., 2012. Nicotine effects on brain function during a visual oddball task: a comparison between conventional and EEG-informed fMRI analysis. J. Cogn. Neurosci. 24, 1682–1694.
Warbrick, T., Mobascher, A., Brinkmeyer, J., Musso, F., 2011. Direction and magnitude of nicotine effects on the fMRI BOLD response are related to nicotine effects on behavioral performance. Psychopharmacology 215, 333–344.
Woolrich, M., Behrens, T., Beckmann, C., 2004. Multilevel linear modelling for FMRI group analysis using Bayesian inference. Neuroimage 21, 1732–1747.
Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., 2002. A general statistical analysis for fMRI data. Neuroimage 15, 1–15.
Yan, C.-G., Craddock, C.R., Zuo, X.-N., Zang, Y.-F., Milham, M.P., 2013. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. Neuroimage 80, 246–262.
Yendiki, A., Greve, D., Wallace, S., Vangel, M., Bockholt, J., 2010. Multi-site characterization of an fMRI working memory paradigm: reliability of activation indices. Neuroimage 53, 119–131.
Zhang, N., Zhu, X., Chen, W., 2005. Influence of gradient acoustic noise on fMRI response in the human visual cortex. Magn. Reson. Med. 54, 258–263.