Supplemental Methods and Materials

Methods

MRI acquisition

A large MRI data set representing normal young adults (N = 1405, mean age 21.4 ± 2.9 years, 57% female) was obtained from the Harvard-MGH Brain Genomics Superstruct Project (1) (Table 1). Imaging was performed with a 3T Siemens Tim Trio scanner equipped with a standard 12-chanel head coil. Functional images were acquired using a gradient echo echo-planar sequence (frame TR 3.0 sec, TE 30 ms, flip angle 85°, $3 \times 3 \times 3$ mm voxels). During rs-fMRI, subjects were instructed to remain still, stay awake, and keep their eyes open. Two rs-fMRI runs (124 frames each) were acquired per subject. Total rs-fMRI time was 12.0 minutes, discounting the first 4 frames of each run. Anatomical imaging included one sagittal T1-weighted magnetization prepared multi-echo (MP–RAGE) scan (T1W) and one T2-weighted scan (T2W). Additional imaging details are given in (2).

fMRI preprocessing

fMRI preprocessing was as described in (3). Briefly, this included compensation for slice-dependent time shifts, elimination of systematic odd-even slice intensity differences due to interleaved acquisition, and rigid body correction of head movement within and across runs. Atlas transformation was achieved by composition of affine transforms connecting the fMRI volumes with the T2W and T1W structural images. Additional preprocessing in preparation for latency analysis included spatial smoothing (6 mm full width at half maximum (FWHM) Gaussian blur in each direction), voxel-wise removal of linear trends over each fMRI run, and temporal low-pass filtering retaining frequencies below 0.1 Hz. Spurious variance was reduced by regression of nuisance waveforms derived from head motion correction and timeseries extracted from regions (of "non-interest") in white matter and CSF as well the BOLD timeseries averaged over the brain (4). Frame censoring was computed at a threshold of 0.5% root mean square frame-to-frame intensity change (5). Epochs containing fewer than 10 contiguous frames were excluded. These criteria removed 3.5 ± 1.3% of frames from the functional connectivity computations. Subjects with less than retained 100 frames were excluded altogether, as were subjects with an outlying T1W→atlas voxel similarity measure. Application of these quality assurance criteria left 1376 usable datasets that were randomly divided into 2 cohorts of 688 subjects each.

Computation of lag between BOLD time series

Our method for computing lags between time series has been previously published (3). We briefly recapitulate the methodology here. Conventional seed-based correlation analysis involves computation of the Pearson correlation, r, between the time series, $x_1(t)$, extracted from a seed region, and a second time series, $x_2(t)$, extracted from some other locus (single voxel or region of interest). Thus,

$$r_{x_1 x_2} = \frac{1}{\sigma_{x_1} \sigma_{x_2}} \frac{1}{T} \int x_1(t) \cdot x_2(t) dt, \qquad [S1]$$

where σ_{x_1} and σ_{x_2} are the temporal standard deviations of signals x_1 and x_2, and T is the interval of integration. Here, we generalize the assumption of exact temporal synchrony and compute lagged cross-covariance functions. Thus,

$$C_{x_1 x_2}(\tau) = \frac{1}{T} \int x_1(t + \tau) \cdot x_2(t) dt, \qquad [S2]$$

where τ is the lag (in units of time). The value of τ at which $C_{x_1 x_2}(\tau)$ exhibits an extremum defines the temporal lag (equivalently, delay) between signals x_1 and x_2 (6). Although cross-covariance functions can exhibit multiple extrema in the analysis of periodic signals, BOLD time series are aperiodic (7, 8), and almost always give rise to lagged cross-covariance functions with a single, well defined extremum, typically in the range ± 1 sec. We determine the extremum abscissa and ordinate using parabolic interpolation (Figure S1).

Given a set of n time series, $\{x_1(t), x_2(t), \cdots, x_n(t)\}$, extracted from n regions of interest (ROIs), a lagged cross-covariance function can be computed between every pair of time series. Thus,

$$C_{x_i x_j}(\tau) = \frac{1}{T} \int x_i(t + \tau) \cdot x_j(t) dt, \qquad i, j \in 1, 2, \dots, n. \qquad [S3]$$

$C_{xx}(\tau)$ is an $n \times n$ matrix that describes the covariance structure of the signal system parametric in lag. Finding all $\tau_{i,j}$ corresponding to the extrema, $a_{i,j}$, of $C_{x_i x_j}(\tau)$ yields the anti-symmetric, time delay matrix:

$$TD = \begin{bmatrix} \tau_{1,1} & \cdots & \tau_{1,n} \\ \vdots & \ddots & \vdots \\ -\tau_{1,n} & \cdots & \tau_{n,n} \end{bmatrix}. \qquad [S4]$$

The diagonal entries of TD are necessarily zero, as any time series has zero lag with itself. Moreover, $\tau_{i,j} = -\tau_{j,i}$, since time series $x_i(t)$ preceding $x_j(t)$ implies that $x_j(t)$ follows $x_i(t)$ by the same interval.

Computation of lag threads from the time delay matrix

The matrix, TD_z, is derived from TD by forcing each column to be zero-mean. Each column indexed by i, $i \in \{1, 2, \dots, n\}$, then is a zero-centered lag map. We next apply principal components analysis (PCA) to TD_z as a method for clustering the lag maps in TD_z. Before proceeding, we note that because PCA is being applied to find clusters in a set of lag maps, it is not necessary that TD_z be square. Here, we let TD_z be $m \times n$, where m (number of rows) $> n$ (number of columns). The rationale for this choice will be explained shortly. PCA is applied to TD_z by computing C, the $n \times n$ covariance matrix of TD_z. Thus,

$$(1/m)[TD_z]^T[TD_z] = C = V \Lambda V^T. \qquad \text{[S5]}$$

The columns of V are the eigenvectors of C; the diagonal entries of Λ contain the eigenvalues. Eigenvalues derived by analysis of real rs-fMRI data are shown in main text Figures 1, 3A, and 4D. Lag thread topographies are computed as eigenvector-weighted sums over lag maps. Thus,

$$L = (1/\sqrt{m})[TD_z]V. \qquad \text{[S6]}$$

The columns of L represent the lag thread topographies in units of time (sec). Lag thread topographies are illustrated in main text Figure 2 and Supplemental Figures S7-S14. Using $V^TV = I$ (the identity matrix), it is a simple matter to show that the columns of L are mutually orthogonal and ordered according to variance explained. Thus,

$$L^T L = V^T \text{CV} = \Lambda, \qquad \text{[S7]}$$

where Λ is the $n \times n$ diagonal matrix of eigenvalues. Since the rank of L generally is much less than n ((9) Fig. 3A), it is useful to truncate L to obtain

$$\tilde{L} = \begin{bmatrix} | & & | \\ | & & | \\ | & & | \\ l_1 & \dots & l_k \\ | & & | \\ | & & | \\ | & & | \end{bmatrix}, \qquad \text{[S8]}$$

where $k < n$. We verify the accuracy of this approach, we applied this procedure to the toy cases shown in Fig. 1 and Figs. S2-S4, animated either with impulse functions (as shown) or "1/f" spectral content duplicated from the BOLD rs-fMRI signal, and found that we were able to correctly recover the forward model.

When the number of ROIs is small, it is most straightforward to define TD as square and skew-symmetric ($TD^T = -TD$). Thus, in the illustrations shown in Fig. 1 and Supplemental Figures S2-S4, $m = n = 6$. However, in real BOLD fMRI data, when the

number of ROIs is large, defining TD as skew-symmetric leads to an under-estimation of the rank of L because of the limited signal-to-noise ratio of BOLD rs-fMRI. Precisely this situation was encountered in our previous report (3), in which TD was defined as 1849×1849 on the basis of (9 mm)^3 ROIs, leading to an estimated lag process rank of 2. Here, we calculate lag maps at (6 mm)^3 voxel resolution using reference signals extracted from (15 mm)^3 cubic ROIs, with the voxels and reference regions both uniformly distributed over gray matter. This strategy represents a compromise between preserved spatial resolution and noise suppression effected by averaging over relatively large reference regions. Thus, the results shown in main text Figures 2-3 were derived using TD matrices defined as 6526×330. These TD matrices necessarily are not skew-symmetric. However, the results shown in main text Figure 4 were obtained using perfectly skew-symmetric TD matrices in which $m = n = 17$. The results in Figure 4 demonstrate that lag processes of rank 8 (or 6, depending on estimator) can be reproducibly extracted from a modest number or ROIs, provided that these ROIs represent terminal (beginning or ending) nodes of lag threads.

Correlation across lag threads

As is evident in [S7], $L^T L = \Lambda$ is $n \times n$. Similarly, $\tilde{L}^T \tilde{L} = \Lambda_k$ is $k \times k$. The key step in relating lag threads to zero-lag temporal correlations is computation of the $m \times m$ (voxels $\times$ voxels) outer product,

$$S = (1/k)\tilde{L}\Lambda_k^{-1}\tilde{L}^T. \qquad\qquad\qquad [S9]$$

In [S9], the factor, Λ_k^{-1}, ensures that the values in S are unit-less and confined to the range [-1, +1]. Numerically, these values are equivalent to Pearson correlations over lag threads. The relation between lag threads and zero-lag temporal correlation is best observed by masking S to include only voxels with at least a 90% chance of belonging to one of the seven resting state networks defined in (10). These voxels correspond to the intersection of the (6 mm)^3 voxels and the RSNs shown in Supplemental Fig. S5. The total number of such voxels is 1065. Hence, in main text Figures 5A and 5B, the matrices are 1065×1065.

Simulating synthetic BOLD fMRI time series

To better understand the connection between lag threads and zero-lag temporal correlation structure (Figure 5), we generated synthetic timeseries, duplicating the spectral content of real BOLD rs-fMRI data. Power spectra of real data were estimated using the fast Fourier transform and averaging over all (6 mm)^3 gray matter voxels and all subjects in the first group of 688. Repeating this procedure in the second group of 688 subjects yielded nearly identical results. Log-log plots confirmed that these spectra were very similar, in the 0.01 Hz to 0.1 Hz frequency band, to previously reported

results (7, 8, 11). Synthetic timeseries then were generated by filtering white noise. We did not attempt to model regional differences in BOLD rs-fMRI spectral content as these differences are minor.

Simulating lag threads

The following is our basic strategy for simulating BOLD signal timeseries constructed from lag threads: the synthetic timeseries were used to realize lag threads with realistic spectral content. Each voxel in a thread was assigned the same timeseries, shifted by the appropriate, thread-determined lag, as illustrated in main text Figure 6B. The shift operation was implemented using sinc interpolation. Each lag thread was independently realized. To obtain the results shown in main text Figures 7B and 8C, multiple lag thread timeseries were temporally concatenated, where the ratios of the lengths of the timeseries corresponding to each lag thread were matched to the ratio of the lag thread eigenvalues. Temporal concatenation was followed by computation of zero lag temporal correlation matrices. Superposition of thread timeseries by averaging generated essentially the same results, which is predictable, as thread realizations are uncorrelated.

In more detail, let $\tilde{L}$ represent the set of 8 lag threads, as in [S8]. For each j, l_j is then simply a vector of lag values, $l_{i,j}$; $i \in \{1,2,\dots,6526\}$, where 6526 is the number of voxels in the system, and j indexes thread number. Let $f_j(t)$ be a one-dimensional, synthetic realization of a BOLD signal. Then an individual lag thread l_j is realized as a multi-dimensional BOLD time series by producing $v_j = f_j(t + l_{i,j})$. The expression v_j is a multi-dimensional time series where each voxel i is assigned the timeseries $f_j(t)$ offset by the appropriate lag, $l_{i,j}$. Repeating this procedure for each lag thread, we create 8 synthetic multi-dimensional BOLD timeseries. We then temporally concatenate these time series to produce the BOLD timeseries, $V = [v_1, \dots, v_8]$. To respect the hierarchy of lag threads, the ratios of the lengths of each v_j are scaled by the their respective lag thread eigenvalues. Finally, to match the analysis of real rs-fMRI data, global signal regression was applied before computing the zero-lag temporal correlation structure of the time series.

Generating orthogonal simulated lag threads

As lag threads in real BOLD rs-fMRI data are computed by PCA, they are orthogonal by definition [S6-S7]. Accordingly, this feature was built into the lag thread simulations (main text Figures 1 and 7). Simulating orthogonal lag threads is trivial in the absence of motifs. The challenge is preserving shared motifs in orthogonal lag sequences. In the simulation shown in main text Fig. 7, there are 30 nodes, 8 lag threads and 2 motifs and both motifs are shared across all lag threads; in real data, all motifs may not be shared

across all lag threads. To simplify the simulation, all threads are modeled as propagating in lock-step (i.e., at unit intervals of time) through all nodes in sequence. (This simplification causes the distribution of inter-node lags to assume the form of a symmetric triangle peaking at 0; in real data this distribution appears to be approximately normal.) The motif constraint fixes the sequence of propagation though nodes 1-5 and nodes 6-10, respectively (see Supplemental Figure S6). The orthogonality condition was enforced by repeatedly permuting the thread sequences in all nodes outside the motifs, i.e., in nodes 11-30. [7] ($L^T L = \Lambda$, here 8×8) was computed in each trial. Perfect orthogonality would correspond to zero in all off-diagonal terms. As an index of orthogonality, we evaluated the ratio of the mean squared off-diagonal terms to mean squared diagonal terms ($(1/28)\sum_{i \neq j} \lambda_{i,j}^2 / (1/8)\sum_i \lambda_{i,i}^2$). Over 10^7 trials, the most nearly orthogonal simulated L (Fig. S6) had an index of orthogonality = 0.06. Thus, in main text Fig. 7, the simulation was approximate.

Dimensionality attributable to a single shared thread motif

In the main text, in connection with Fig. 8A, we note that the dimensionality of intra-RSN lag structure should be 1, if RSNs correspond to lag thread motifs, i.e., common sequences of propagation shared over lag threads. Implicit in this logic is that nodes belonging to a single motif should have a dimensionality of 1. Equivalently, the first eigenvector of intra-motif lag structure should account for 100% of the variance. In real data, in the brain as a whole, the first eigenvector accounts for 14% of the variance (main text Fig. 3A). In analyses restricted to within-RSN voxels (Fig. 8A), the first eigenvector accounted for 45% - 70% of the variance, excluding the SMN (40% in the SMN).

The connection between RSNs and lag thread motifs is illustrated in main text Figure 7. The lag structure of all 30 nodes taken together has a dimensionality of 8, because there are 8 threads. The first eigenvector accounts for 19% of the variance. However, nodes 1-5 and nodes 6-10 constitute two motifs, i.e., "RSNs" (main text Fig. 7B). The lag structure restricted to a single motif, i.e., nodes 1-5 (or nodes 6-10) contains only one thread; the first eigenvector accounts for 100% of the variance and the dimensionality is 1, as in the case illustrated in Supplemental Figure S3. Thus, evidence of the connection between lag thread motifs and RSNs is low dimensionality of the lag system, or, equivalently, the high proportion of variance in the first eigenvector.

Zero-lag temporal correlation need not specify lag structure

Fig. 8 demonstrates that lag structure can give rise to zero-lag temporal correlations. The reverse is not true. That is, zero-lag temporal correlations need not impose any lag thread structure. This lack of relation is demonstrated in Fig. 9, in which we produce synthetic BOLD timeseries with zero-lag temporal correlation structure matched to real

data. We describe here our method for producing synthetic BOLD timeseries with a specific correlation structure. Let C be the covariance structure derived from real rs-fMRI data (averaged over 688 subjects). In the present application, C is a 36 × 36 matrix (12). We compute the eigen-decomposition of $C = W\Lambda W^t$. Now, suppose Z is a matrix of random normal deviates with dimensions ROIs × time (in this case, 36 × 2048 seconds). The timeseries defined as $F = Z\Lambda W^t$ then has covariance structure matched to C. Finally, we filter the timeseries in F, as described previously, to produce the timeseries B with spectral content matched to the real rs-fMRI BOLD signal.

Lagged cross-covariance curves exhibit a single peak

In general, it is possible for lagged-cross covariance (CCV) curves to exhibit multiple peaks, in which case choosing only a single peak to define a lag between regions would represent an oversimplification of the relevant dynamics. One obvious driver of multiple peaks in lagged CCV curves would be periodicity in the signals; however, we have previously noted that the rs-fMRI BOLD signal are known to be aperiodic (7, 13). Another mechanism that could give rise to multiple lagged CCV curve peaks are looped or re-entrant temporal sequences. However, in the case of the rs-fMRI BOLD signal, we believe it is valid to consider only single peaks in lagged CCV curves. We have examined hundreds of BOLD timeseries cross-covariance curves derived from ROI pairs of interest. In effectively all cases, we observe only a single peak in the range ±5 sec (Figs. S25A-S25D). Secondary local extrema are rarely observed, and always appear at lag magnitudes > 6 sec, and always are of comparatively low magnitude (Fig. S25E). It is formally possible that these rare secondary peaks represent evidence for re-entrant temporal architectures as illustrated in Fig. S25F. To address reentrancy, we simulated the 3-node reentrant system in Fig. S25F realized using time-series with spectral content duplicated from real BOLD data. The simulation has a lag of one unit between each node in the system. The resulting lagged CCV curve (Fig. S25G) shows 2 peaks in the relation between nodes 1 and 3. The first peak is at -2, as node 1 leads node 3 by 2 units, and the second peak is at +1, as node 3 leads node 1 by one lag unit. Thus, reentrancy manifests as lagged CCV functions with multiple peaks of comparable magnitude. In the absence of consistent evidence of such lagged CCV curves in our data, we believe it unlikely that reentrant loops are a major feature of BOLD rs-fMRI lag structure. Additionally, although the coupling dynamics relating neural activity to the BOLD signal ("hemodynamic impulse response function") are incompletely understood mechanistically, it is clear that the spectral content of BOLD event-related responses is markedly attenuated above ~0.2 Hz, and there is no reason to suppose that it is otherwise in resting state activity (14, 15). In view of these considerations, we believe that it is valid to measure lags in the ±5 sec range by parabolic interpolation of CCV curves and to ignore secondary peaks at "extreme" lags. In effect, we are concerned only with temporal lags in infra-slow phenomenology. Within

this scope, we believe that our strategy does not simplify away or distort the essential phenomenology.

<u>Averaging to compute group level lag structure</u>

We compute TD_z matrices in individuals, and then average these matrices to produce a group-level TD_z. Alternatively, we could have first averaged lagged CCV curves over subjects and then applied parabolic interpolation to produce a group level TD_z matrix. In either case, PCA would be applied to the group-level TD_z matrix. The reason we do not take the second approach is that it is computationally prohibitive, in large datasets, to write each subject's set of cross-covariance curves to memory. However, to address whether both approaches yield the same answer, we compared the two strategies for computing group-level $17 \times 17\ TD_z$ matrices. The result, shown in Fig. S26, demonstrate that the TD_z estimates are nearly identical.

<u>Group size required to compute lag structure</u>

In the main text, lag structure is computed at the group level in two very large, independent populations (N = 688 per group). This raises the question of how many subjects are required to observe lag threads at the group level. Regarding this question, it is useful recall that TD matrices are derived from lagged cross-covariance (CCV) analyses of high-dimensional timeseries. As lagged CCV curves are second order statistics, CCV measurement is affected by sampling error, the variance of which is inversely proportional to the number of independent samples. Owing to temporal auto-correlation, BOLD fMRI yields one independent sample every ~6 seconds. The present analyses are based on 720 sec of fMRI data per subject (discounting censored frames), i.e., ~120 independent CCV samples/subject. Expressed in terms of standard errors, the precision of each CCV estimate then is on the order of $\pm 1/\sqrt{120}$, or ±9%. Lag estimates also are affected by instrumental errors. Important sources of instrumental error include the fact that lags must be estimated by interpolation of sparsely sampled fMRI data (1 frame every 3 sec). Electronic noise is another important source of instrumental error. Electronic noise depends on acquisition ROI size (16). Suppression of electronic noise by extracting reference signals from large (15mm cubic) ROIs represents an important component of the present strategy for computing 6526×330 TD matrices.

The preceding considerations pertain to measuring lags over ROI pairs in individuals. However, lag threads are derived by eigenanalysis of TD_z matrices averaged over individuals. Deriving error bounds on these quantities from first principles would be exceedingly difficult. We therefore, adopt an empirical approach to this question. Fig. S27 plots convergence of TD matrix estimates depending on number included subjects. To make these estimates, we calculated two quantities in the first group of 688 subjects:

(I) the average TD matrix in our 17×17 ROI system (as in main text Fig. 4), and (II) the average TD matrix in our 6526×330 ROI system. We then computed both of these quantities in sub-samples of the second, independent group of 688 subjects and plotted the correlation (averaged over unique matrix cells) between the estimates. The 17×17 ROI system achieved excellent convergence after 50-100 subjects. In contrast, more than 400 subjects were required to achieve satisfactory convergence of the 6526×330 ROI system. This difference reflects variable suppression of electronic noise in the two cases, depending on ROI size: $(6mm)^3$:$(15mm)^3$ cubes in the 6526×330 case vs. ~7191mm^3:7191mm^3 irregular regions in the 17×17 case (ROI voxel count range 54 to 763, mean = 266). Thus, it takes more data to reconstruct the topography of threads than it does to estimate the dimensionality of the lag system on the basis of *a priori* selected ROIs. We note that the analysis illustrated in Fig. S27 does not distinguish between imprecision in the measurement of lags vs. inter-individual variability in lag structure. It is likely that the number of subjects required to achieve TD matrix convergence would be less if the quantity of data obtained in each subject were increased. However, the precise outlines of this trade-off are not known.

<u>Validity of applying PCA to recover lag thread topographies</u>

It is reasonable to ask whether PCA is an effective tool for decomposing the zero-centered time delay matrix into a set of lag maps. To provide an answer, it is worth clearly stating the role of PCA in our analysis. Specifically, our application of PCA can be thought of as a clustering algorithm. As noted in the main text, when we compute seed-based lag maps using various loci as seeds, we often observe incompatible lag patterns. In other words, in one seed-based lag map, region A may be earlier than region B, whereas the situation is reversed in a second seed-based lag map; the anterior insula vs. posterior insula seeded lag maps (Figs. S15 and S16) illustrate this point. However, other seed-based lag maps are quite similar, e.g., the putamen and entorhinal cortex seed-based lag maps (Figs. S18 and S19). Given a set of 330 lag maps (columns of the matrix TD_z) we wish to identify all sets of compatible lag maps, e.g., those seeded from the putamen and entorhinal cortex. The PCA-derived topographies approximately identify the full set of unique, non-compatible lag systems (i.e., "lag threads") that are superposed in BOLD rs-fMRI data.

PCA is a linear technique that imposes orthogonality in choosing lag threads. Three pieces of evidence suggest that our results are not markedly dependent on the assumptions of linearity and orthogonality. First, to explore whether linearity significantly determines the computed topography of lag threads, we applied kernel PCA to TD_z (using a Gaussian kernel) and obtained topographies (Fig. S28) very similar to those shown in main text Fig. 2. This result suggests that lag thread topographies are robust

to alternative clustering strategies. Second, to test whether lag threads obtained by linear PCA are representative of seed-based lag maps, we computed spatial correlations between lag thread topographies (T1-T8) and the topographies of 17 seed-based lag maps (Fig. S29). This result shows that similar seed-based lag maps are represented together in a single lag thread. For instance, seed-based lag maps computed from the PCC, putamen, and entorhinal cortex, all of which appear quite similar by eye (see Figs. S18, S19, and S21), are strongly correlated with lag thread 1. Moreover, incompatible seed-based lag maps, such as anterior insula and posterior insula, are strongly represented in different lag threads: thread 2 vs. thread 3 in this example. The separation of seed-based lag maps is not perfect as lag threads were generated by applying PCA to lag maps generated from 330 lag maps, not just the 17 maps used here as references. However, even in this imperfect demonstration, at least the first four lag threads separate these 17 lag maps derived from the whole brain well enough to suggest that the topographies of the first four lag threads are meaningful as reported. As a side note, it is noteworthy that the lag thread 5, which likely captures venous outflow, is only weakly correlated with each of the seed-based maps.

Finally, to further demonstrate that lag threads are a complete decomposition of time delay structure in BOLD rs-fMRI, we computed the eigenvalue-weighted sum of the 8 lag threads reported in the manuscript (Fig. S30B). For comparison, we show the lag projection map reported in our previous publication (Fig. S30A; (3)), which was computed by taking the column-wise mean of the ROI x ROI time delay matrix. As is evident, the two maps are very similar (spatial correlation r = 0.71). Therefore, we believe that the topographies of the lag threads derived by PCA represent a reasonably accurate decomposition of rs-fMRI time delay structure. The correspondence between the lag projection (Fig. S30A) and weighted lag thread sum (Fig. S30B) also constitutes evidence that the sign of the lag thread maps were correctly chosen.

Supplemental Figure Captions
Figure S1: Calculation of pair-wise timeseries lags using cross-covariance and parabolic interpolation. Lags are defined by analysis of timeseries derived from two loci. A: Two exemplar loci (both in the default mode network). B: Time series sampled over ~300 s extracted from the illustrated loci. C: The corresponding lagged cross-covariance function, equation [S2]. The lagged cross-covariance is defined over the range $\pm T$, where T is the run duration. The range of the plotted values is restricted to ±12 s, which is equivalent to ±4 frames (red markers) when the repetition time is 3s. The lag between the time series is the value at which the [absolute value of the] cross-covariance function is maximal. D: This extremum can be determined at a resolution finer than the temporal sampling density (one frame every 3 seconds) by parabolic interpolation (green line) through the computed values (red markers). This extremum (arrow, yellow marker) defines both the lag between time series i and j ($\tau_{i,j}$; equation [S4]) and the corresponding amplitude ($a_{i,j}$; equation [S4]). Panels A and B are adapted from (17). A

systematic lag between the two timeseries is visually apparent. Panels C and D are adapted from (3).

Figure S2: Illustration of a perfectly synchronous system. This figure is formatted identically to Fig. 1. Panel A shows 6 timeseries consisting of perfectly synchronous timeseries. Panel B illustrates the absence of propagation. Panel C shows a time delay matrix (TD), equation [S4], that is identically zero. Panel D shows the eigenvalues of TD_z, equation [S5], which are identically zero.

Figure S3: Illustration of a system with one lag thread. This figure is formatted identically to Figs. 1 and S2. Panel A shows 6 time series consisting of lagged timeseries. Panel B illustrates the propagation through nodes. Panel C shows the time delay matrix (TD). All rows are the same except for an additive constant. The lag structure of this system is fully captured in the column-wise mean projection, as shown in the bottom row of (C). Panel D shows the eigenvalues of TD_z, (equation [S5] in SI). As there is exactly one lag thread, there is exactly one non-zero eigenvalue.

　　　The same algebra applies to lag thread motifs, i.e., shared sequences of propagation across multiple lag threads within restricted sets of voxels. Thus, the dimensionality of the lag structure within an ideal thread motif is exactly 1, as in panel D. Equivalently, all (100%) of the lag variance within an ideal thread motif should be captured in the first eigenvector. The thread motif model posits that zero lag correlations within RSNs arise on the basis of shared motifs. As shown in Fig. 8, the *maximum likelihood* dimensionality of lag structure within a priori defined RSNs is indeed 1, in 7 out of 8 cases (the exception being the SMN). However, in real data, the proportion of lag variance within RSNs captured by the first eigenvector is only 70% to 45%. Thus, the motif model accounts for RSNs as defined in (10), albeit imperfectly.

Figure S4: Illustration of a system with one lag thread. This figure is identical Fig. S3 except the sign of half the timeseries is inverted. The zero-lag correlation structure of the systems in Figs. S3 and S4 differ but the lag structure is the same. This example serves as a simple demonstration that zero-lag temporal correlation structure need not uniquely determine lag structure.

Figure S5: Gray matter voxels and regions of interest. Lag maps and lag threads in Figs. 2-3 are based on a parcellation of gray matter into 6526 (6 mm)^3 voxels, represented in panel A. Panel B shows a parcellation of gray matter into 330 (15 mm)^3 cubic regions of interest (equivalently, "large voxels"). These parcellations were used to create the rectangular, 6526 × 300 time delay matrices submitted to principal components analysis (equation [S5]). Panel C shows the 1065 (6 mm)^3 gray matter voxels satisfying the criterion of >= 90% probability of belonging to one of the 7 resting state networks defined in (10). This subset of voxels appears in the correlation matrices shown in Figs. 5A, 5B, and 7C. The original parcellation (10) lumped together the auditory network (AUD) and the sensorimotor network (SMN). Here, we have defined these networks as distinct, thereby increasing the total number of RSNs to 8. As in Fig. 4, the color code in this figure has nothing to do with latency.

Figure S6: Thread motifs. Panels A and B recapitulate Figure 6. Panel C shows an alternative coding of the information in panel A in which voxel identities are listed according to their order in the thread sequence. Motifs are defined as common orderings across lag threads. The threads are very simple in Fig. 6 but not simple in Fig. 7. Panel D shows the voxel ordering within threads used to create the simulation shown in Fig. 7. There are two thread motifs in this simulation, highlighted by the red and blue boxes. The voxel sequence within these motifs is preserved across all lag threads.

Figures S7-S14: Lag thread maps corresponding to the first 8 eigenvectors (equation [S8]) obtained in the first group of 688 subjects. Note generally symmetric topographies. The corresponding eigenvalue plot is shown in Fig. 3A. Although lag threads superficially resemble RSNs, the relation between these two systems of maps is complex. There is no systematic relation between lag and conventional functional connectivity over voxel pairs (Fig. 5). We hypothesize that RSNs arise as consequence of shared motifs across lag threads (Figs. 7 and 8). Most features of lag threads appear to reflect brain organization. However, Lag Thread 5 prominently includes late voxels near large venous structures, e.g., the superior sagittal sinus, that most likely reflect non-neural mechanisms. Lateness in the venous outflow of the brain, as manifested in lag projection maps, is discussed in (3).

Figures S15-S24: Lag maps calculated using reference time series extracted from selected ROIs in Fig. 4A. Each of these maps demonstrates features of particular interest. In any particular map, the seed structure must appear with a lag value of zero. Therefore, the lag map corresponding to a ROI that is predominantly early, e.g., the anterior insula, must be dominated by late values, as in Fig. S15. Non-transitivity of lag structure is apparent on comparison of multiple maps. For example, in the anterior insula map (Fig. S15), primary sensorimotor cortex is late with respect to anterior insula whereas, in the posterior insula map (Fig. S16), primary sensorimotor cortex is early with respect to anterior insula. A second interesting feature is evident on comparison of the caudate (Fig. S17) vs. the putamen (Fig. S18) lag maps. Most of cerebral cortex is markedly early with respect to the caudate but late with respect to the putamen (lateral cerebellum is late in both maps). Moreover, the sharp contrast between the caudate vs. putamen lag maps demonstrates the existence of sharp boundaries in lag structure. Third, many topographic features are evident in both the seed-based lag maps and the thread maps. This result lends face validity to the PCA strategy because orthogonalization of seed-based maps preserves the underlying structure. For, example, the caudate vs. putamen contrast is clearly evident in Lag Threads 1 (Fig. S7) and Lag Threads 4 (Fig. S10). It may also be noted that Lag Thread 1 (Fig. S7) captures many of the features present in lag maps generated from putamen (Fig. S18), entorhinal cortex Fig. S19), and posterior cingulate precuneus (Fig. S21). Finally, as mentioned in the Discussion, the lag between entorhinal cortex and cortex more generally depends on locus (Fig. S19). Primary sensory areas of cortex (vision, primary somatosensory) are early with respect to entorhinal cortex, whereas higher order cortical areas (medial prefrontal cortex, posterior cingulate precuneus) are late.

Figure S25: Five representative lagged cross-covariance (CCV) curves, computed at the group level in 688 subjects, between pairs of 17 regions of interest (Fig. 4). The x-axes are in units of TR, so one unit represents 3 seconds. Thus, these curves span ±12 seconds. Panels A-D are examples of typical lagged CCV curves, which exhibit a clear single peak. Panel E is an example of a rare lagged CCV curve which has a secondary local extrema at a lag of approximately -7 seconds. Panel F shows a schematic example of a looped or re-entrant architecture. We simulated the system represented in panel F with synthetic timeseries with spectral content duplicated from real BOLD signals. The resulting lagged CCV curve between nodes 1 and 3 is exhibited in panel G and shows 2 clear peaks of similar magnitude. The first peak is at -2, as node 1 leads node 3 by 2 units, and the second peak is at +1, as node 3 leads node 1 by one lag unit. Thus, reentrancy manifests as lagged CCV functions with multiple peaks of comparable magnitude. Multiple peaks are not generally observed in lagged CCV curves computed in real BOLD rs-fMRI data.

Figure S26: Two approaches to group averaging of lag structure yield nearly identical results. We compute TD_z matrices in individuals, and then average these matrices to produce a group-level TD_z. Alternatively, we could have first averaged lagged cross-covariance curves over subjects and then applied parabolic interpolation to produce a group level TD_z matrix. We used the 17 ROIs defined in main text Fig. 4 to compare the TD matrix computed by both strategies. The scatter plot demonstrates that the lag values produced by the two strategies are highly concordant.

Figure S27: Empirical analysis of the number of subjects required to observe lag thread structure at two spatial resolutions. We computed an average TD matrix in our 17×17 ROI system and our 6526×330 ROI system over the first group of 688 subjects. We then computed how correlated these structures are for subsets of a second, independent group of 688 subjects. We show subsets of sizes 1, 10, 20, 50, 100, 200, 400, and 600. We randomly sampled each subset at least 50 times. In the 17×17 case (panel A), which uses a small set of very large ROIs, good convergence is seen already in 50 subjects. In the 6526×330 case (panel B), which entails higher spatial resolution, 200-400 subjects are needed to achieve good convergence. Importantly, convergence is seen in both cases even though the convergent data set is independent of the dataset from which the mean time delay structure is computed. Therefore, these results also further establish the reproducibility of temporal delay structure in BOLD rs-fMRI.

Figure S28: Topographies of the first four components derived by kernel PCA (KPCA) decomposition of the 6526×330 TD_z matrix computed in group 1 (688 subjects). KPCA was applied with a Gaussian kernel with sigma = 1 second (18). The lag thread topographies produced by KPCA, which is a non-linear technique, are highly concordant with those produced by ordinary linear PCA (main text Fig. 2).

Figure S29: Spatial correlations between 8 lag threads (T1-T8; Fig. S7-S14) and seed-based lag maps computed with reference to 17 ROIs defined in main text Fig. 4A. The spatial correlation matrix demonstrates that seed-based lag maps that appear visually compatible are represented strongly in the same lag thread. For example, the putamen

and entorhinal cortex lag maps are both strongly represented in lag thread 1. In contrast, seed-based lag maps that produce incompatible features are represented in different lag threads. For example, the posterior insula lag map is strongly represented in lag thread 2 but not in lag thread 3, whereas the situation is reversed for the anterior insula lag map. Thus, at least the first four PCA derived lag thread topographies reasonably separate features produced by groups of seed-based lag maps.

Figure S30: Lag projection topography is well recovered by computing the weighted sum of lag thread topographies. Panel A shows the lag projection map reported in our previous publication (Fig. S30A; (3)), which was computed by taking the column-wise mean of the ROI x ROI time delay matrix. Panel B shows the eigenvalue-weighted sum of the 8 lag threads reported in the manuscript. As is evident, the two maps are very similar (spatial correlation r = 0.71). Therefore, PCA-derived lag thread topographies represent a reasonably complete decomposition of rs-fMRI time delay structure. The correspondence between the lag projection (A) and weighted lag thread sum (B) also constitutes evidence that the sign of the lag thread maps were correctly chosen.

Movie 1: In order to animate the propagation of lag threads through the brain, the first four lag threads were converted to movies. Each movie consists of 150 frames, where each frame represents a lag between -0.8 and 0.6 seconds. For each frame, the voxelwise lags for each thread were put through a normal probability distribution function with a mean equal to the lag for that frame and a standard deviation of 0.1. This procedure temporally smoothed each lag map by giving voxels high weights in frames with similar lag values. The resulting 4D image was transformed to 1 mm MNI152 atlas space and then spatially smoothed with a 2mm FWHM Gaussian kernel. Movie 1 is an animation of lag thread 1 in Main Text Fig. 2.

Movie 2: Movie 2 is an animation of lag thread 2 in Main Text Fig. 2.

Movie 3: Movie 3 is an animation of lag thread 3 in Main Text Fig. 2.

Movie 4: Movie 4 is an animation of lag thread 4 in Main Text Fig. 2.

Supplemental references

1. Buckner RL, Roffman JL, & Smoller JW (2014) Brain Genomics Superstruct Project (GSP).
2. Thomas Yeo BT, *et al.* (2011) The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* 106(3):1125-1165.
3. Mitra A, Snyder AZ, Hacker CD, & Raichle ME (2014) Lag structure in resting state fMRI. *J Neurophysiol*.
4. Fox MD, Zhang D, Snyder AZ, & Raichle ME (2009) The global signal and observed anticorrelated resting state brain networks. *J Neurophysiol* 101(6):3270-3283.
5. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, & Petersen SE (2012) Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* 59(3):2142-2154.
6. Konig P (1994) A method for the quantification of synchrony and oscillatory properties of neuronal activity. *J Neurosci Methods* 54(1):31-37.
7. He BJ, Zempel JM, Snyder AZ, & Raichle ME (2010) The temporal structures and functional significance of scale-free brain activity. *Neuron* 66(3):353-369.
8. Maxim V, *et al.* (2005) Fractional Gaussian noise, functional MRI and Alzheimer's disease. *Neuroimage* 25(1):141-158.
9. Minka TP (2001) Automatic choice of dimensionality for PCA. *Advances in Neural Information Processing Systems 13*, (MIT Press, Cambridge, MA), pp 598-604.
10. Hacker CD, *et al.* (2013) Resting state network estimation in individual subjects. *Neuroimage* 82C:616-633.
11. He BJ & Raichle ME (2009) The fMRI signal, slow cortical potential and consciousness. *Trends Cogn Sci* 13(7):302-309.
12. Brier MR, *et al.* (2012) Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression. *J Neurosci* 32(26):8890-8899.
13. Adelsberger H, Garaschuk O, & Konnerth A (2005) Cortical calcium waves in resting newborn mice. *Nat Neurosci* 8(8):988-990.
14. Anderson JS (2008) Origin of synchronized low-frequency blood oxygen level-dependent fluctuations in the primary visual cortex. *AJNR Am J Neuroradiol* 29(9):1722-1729.
15. Hathout GM, *et al.* (1994) MR imaging signal response to sustained stimulation in human visual cortex. *J Magn Reson Imaging* 4(4):537-543.
16. Bodurka J, Ye F, Petridou N, Murphy K, & Bandettini PA (2007) Mapping the MRI voxel volume in which thermal noise matches physiological noise--implications for fMRI. *Neuroimage* 34(2):542-549.
17. Fox MD, *et al.* (2005) The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc Natl Acad Sci U S A* 102(27):9673-9678.
18. van der Maaten LJ, Postma EO, & van den Herik HJ (2009) Dimensionality reduction: A comparative review. *Journal of Machine Learning Research* 10(1-41):66-71.

A   Time-course diagram

NODE 1

NODE 2

NODE 3

NODE 4

NODE 5

NODE 6

B   Node Diagram

① 1   ⑥ 6

② 2   ⑤ 5

③ 3   ④ 4

C   TD Matrix

	NODE 1	NODE 2	NODE 3	NODE 4	NODE 5	NODE 6
NODE 1	0	0	0	0	0	0
NODE 2	0	0	0	0	0	0
NODE 3	0	0	0	0	0	0
NODE 4	0	0	0	0	0	0
NODE 5	0	0	0	0	0	0
NODE 6	0	0	0	0	0	0

D   Eigenspectrum
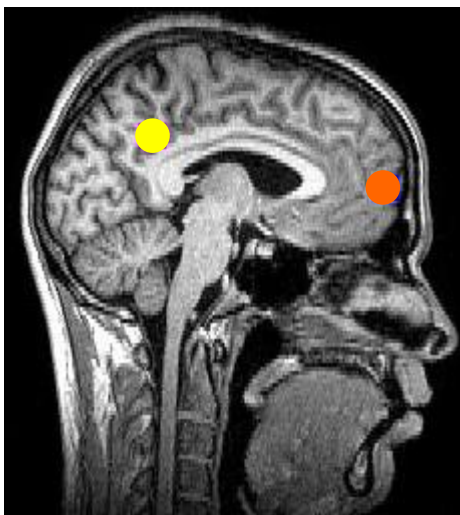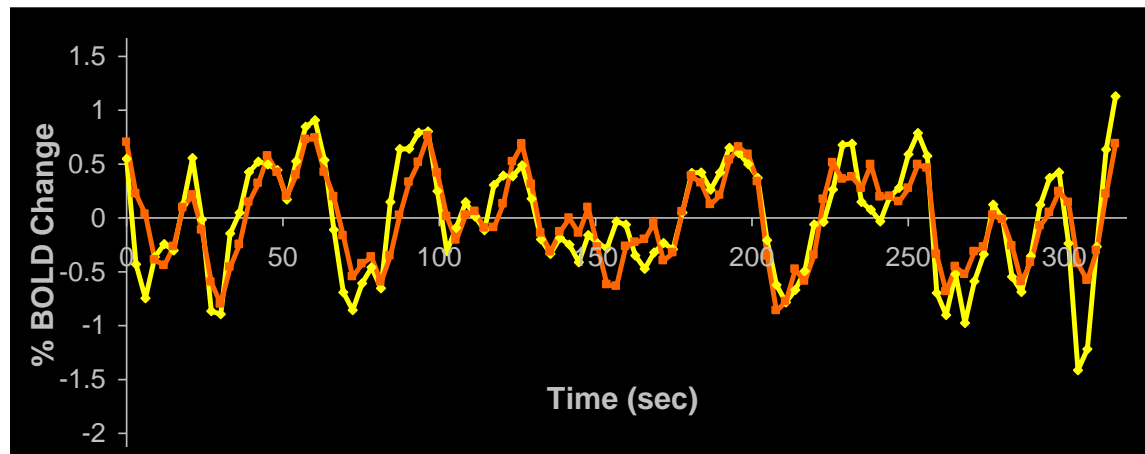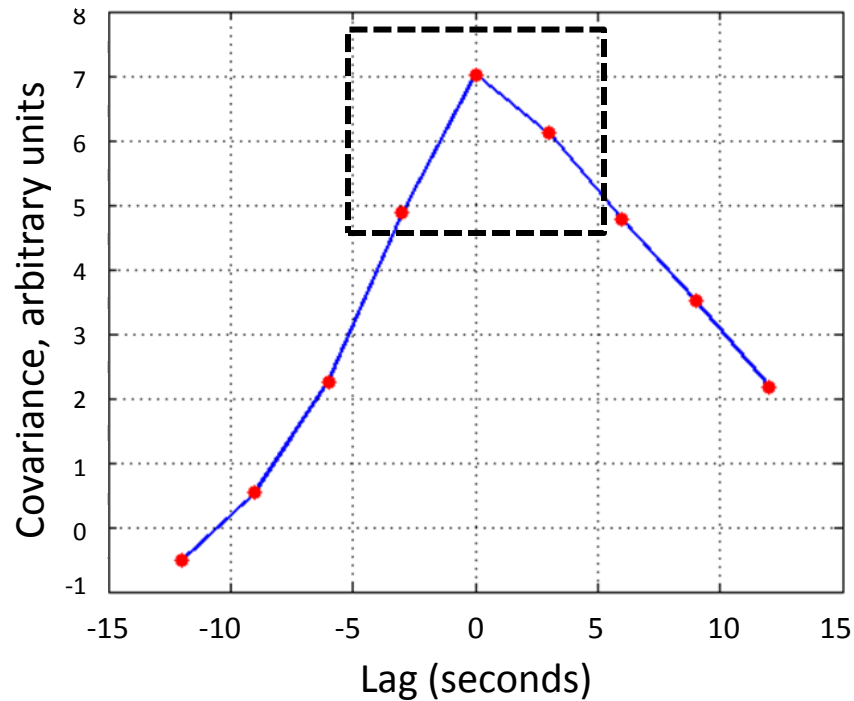
Time-course diagram
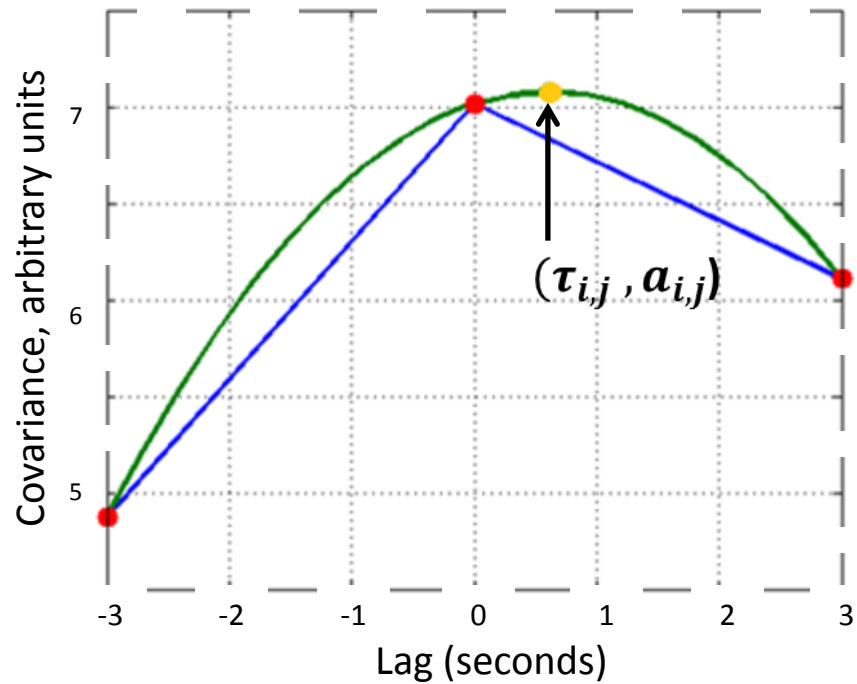
A Time-course diagram
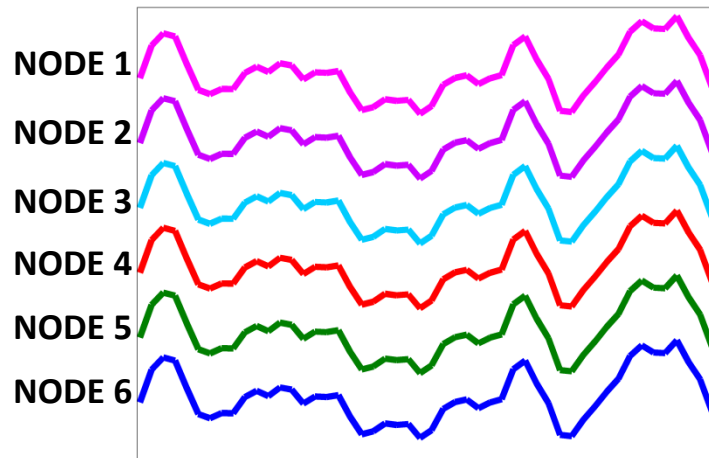
B Node Diagram
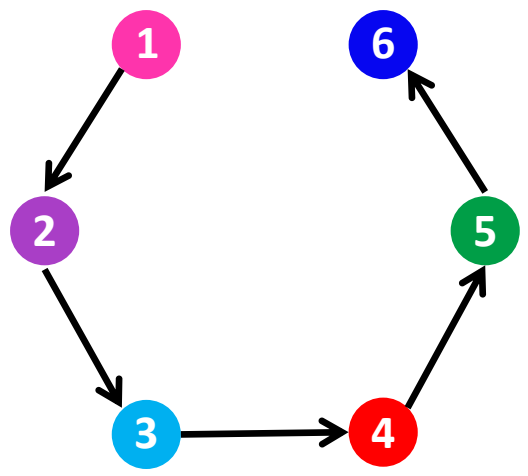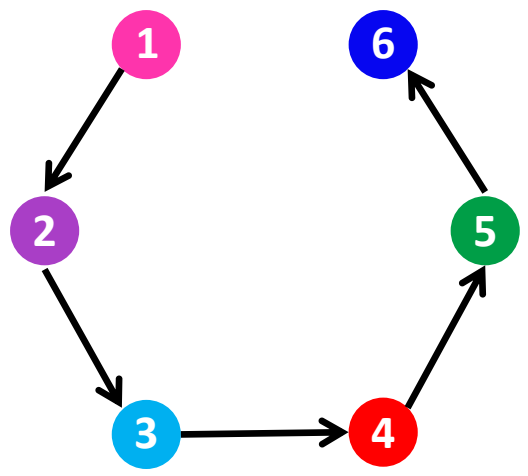
C TD Matrix

	NODE 1	NODE 2	NODE 3	NODE 4	NODE 5	NODE 6
NODE 1	0	1	2	3	4	5
NODE 2	-1	0	1	2	3	4
NODE 3	-2	-1	0	1	2	3
NODE 4	-3	-2	-1	0	1	2
NODE 5	-4	-3	-2	-1	0	1
NODE 6	-5	-4	-3	-2	-2	0
Average	-2.5	-1.5	-0.5	0.5	1.5	2.5

D Eigenspectrum

Time-course diagram

A

Node Diagram

B

TD Matrix

C

	NODE 1	NODE 2	NODE 3	NODE 4	NODE 5	NODE 6
NODE 1	0	1	2	3	4	5
NODE 2	-1	0	1	2	3	4
NODE 3	-2	-1	0	1	2	3
NODE 4	-3	-2	-1	0	1	2
NODE 5	-4	-3	-2	-1	0	1
NODE 6	-5	-4	-3	-2	-2	0
Average	-2.5	-1.5	-0.5	0.5	1.5	2.5

Eigenspectrum

D
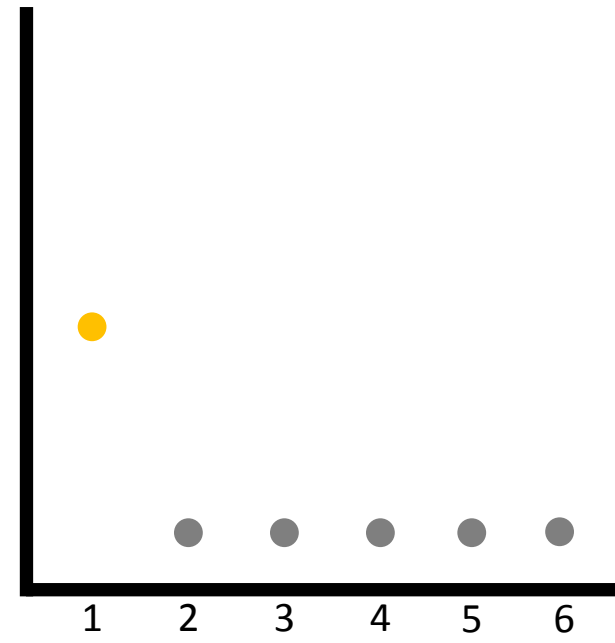
A 6 mm³ voxels
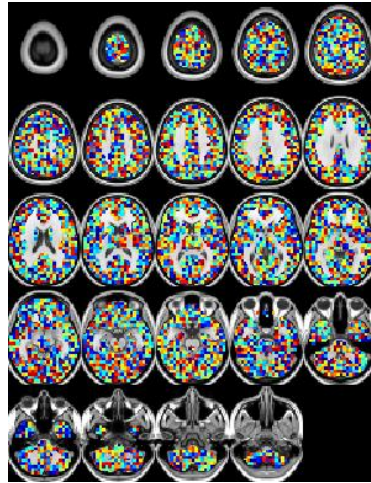
B 15 mm³ ROIs
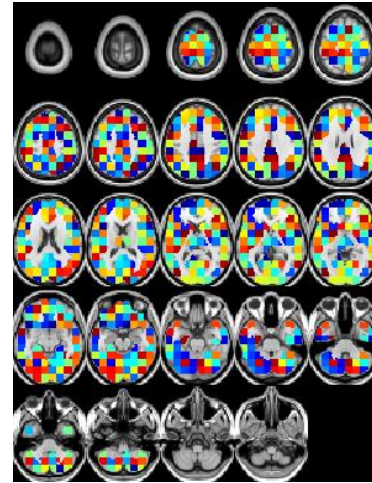
C Resting state network membership
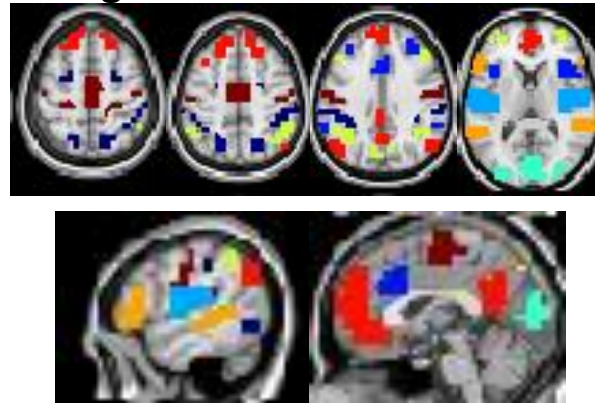
- ● Dorsal attention network (DAN)
- ● Ventral attention network (VAN)
- ● Auditory network (AUD)
- ● Visual network (VIS)
- ● Frontoparietal control network (FPC)
- ● Language network (LAN)
- ● Default mode network (DMN)
- ● Somatomotor network (SMN)

A Thread 1, Thread 2

B Thread 1, Thread 2
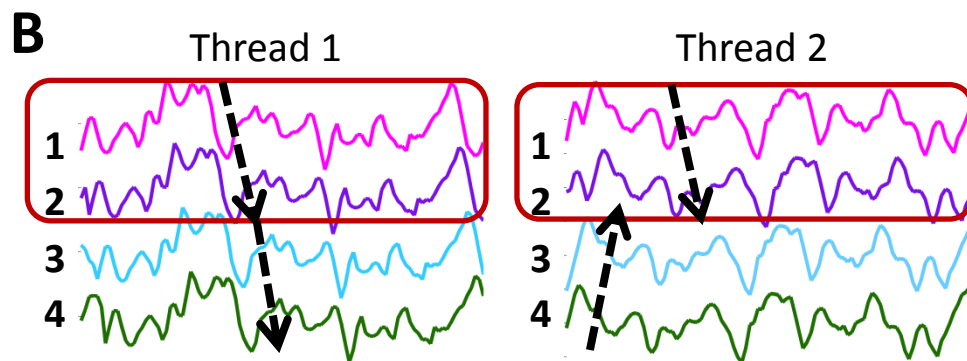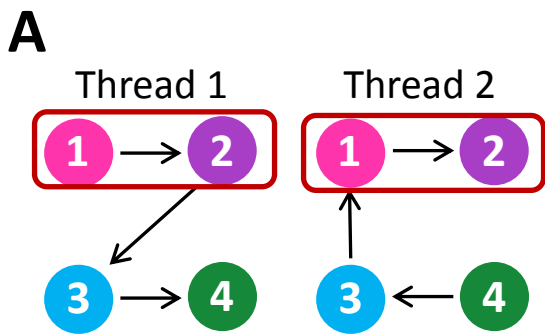
C

voxel #	T1	T2
1	1	3
2	2	4
3	3	2
4	4	1

D

voxel #	T1	T2	T3	T4	T5	T6	T7	T8
1	1	6	25	2	15	20	12	18
2	2	7	26	3	16	21	13	19
	3	8	27	4	17	22	14	20
	4	9	28	5	18	23	15	21
•	5	10	29	6	19	24	16	22
	10	5	15	11	30	12	7	9
	9	4	14	10	29	11	6	8
•	8	3	13	9	28	10	5	7
	7	2	12	8	27	9	4	6
•	6	1	11	7	26	8	3	5
	22	19	9	18	8	14	2	14
•	13	21	2	26	21	29	11	27
	14	26	30	27	10	5	20	12
	21	12	24	28	3	13	19	23
•	30	20	1	21	5	27	8	29
	12	25	4	20	12	28	29	4
•	17	22	10	15	9	18	21	3
	29	23	23	16	14	25	9	13
	27	28	20	29	24	17	10	30
•	28	14	18	17	20	19	24	2
	25	29	22	25	23	15	18	11
•	19	11	8	30	11	26	25	28
	16	18	19	19	6	16	22	15
	11	16	3	24	4	1	27	17
•	20	13	17	12	13	4	28	25
	23	30	7	1	7	2	23	24
•	18	15	5	22	22	3	26	26
	24	27	21	13	25	7	30	16
29	26	17	6	23	2	6	1	10
30	15	24	16	14	1	30	17	1

Thread 1

Thread 2

Thread 3
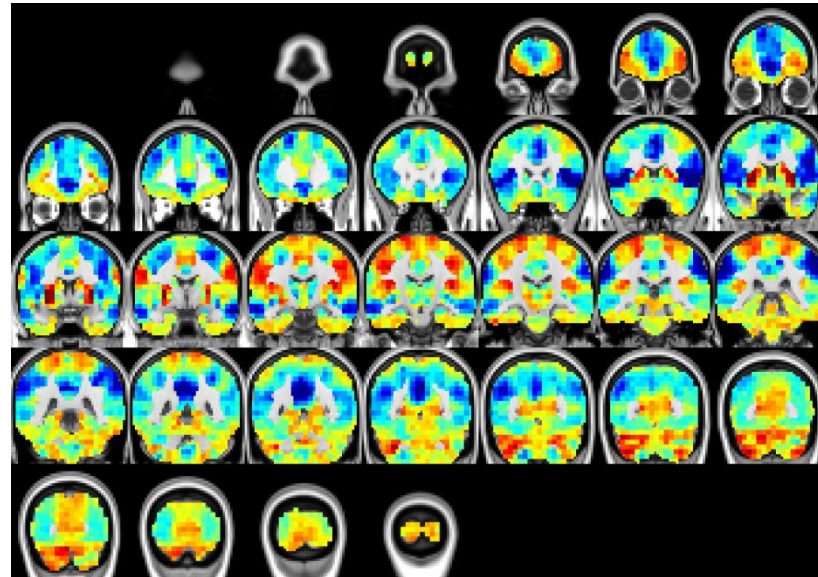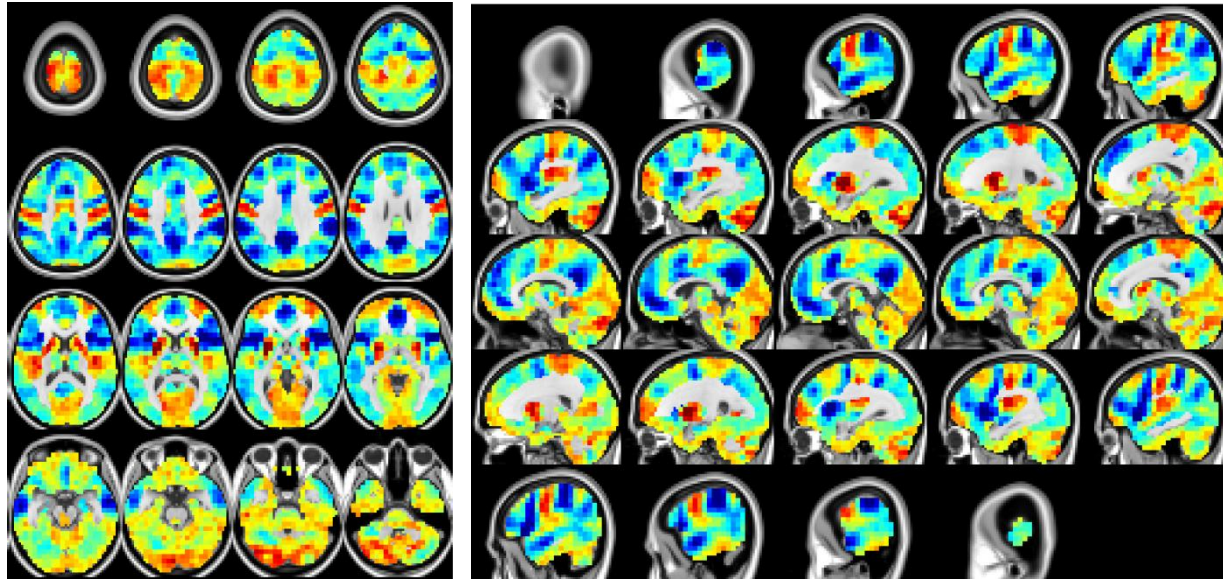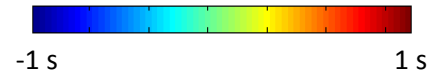
Thread 4
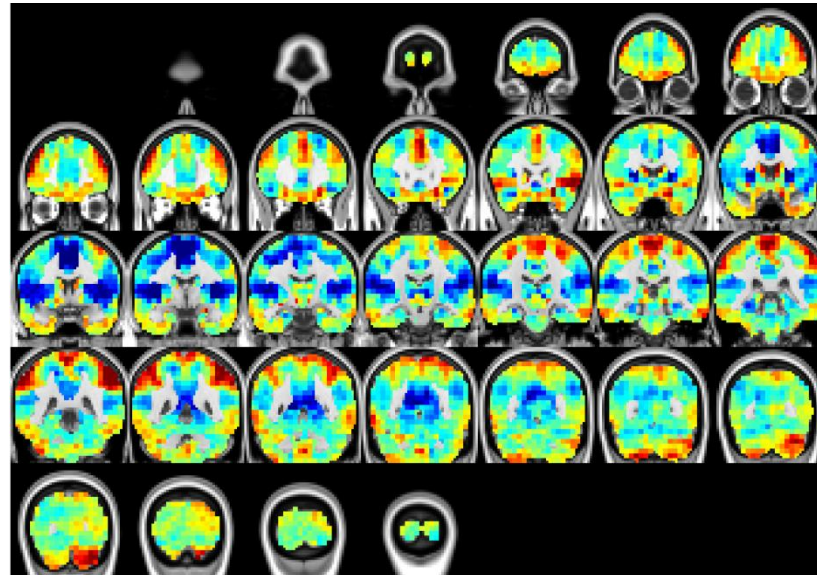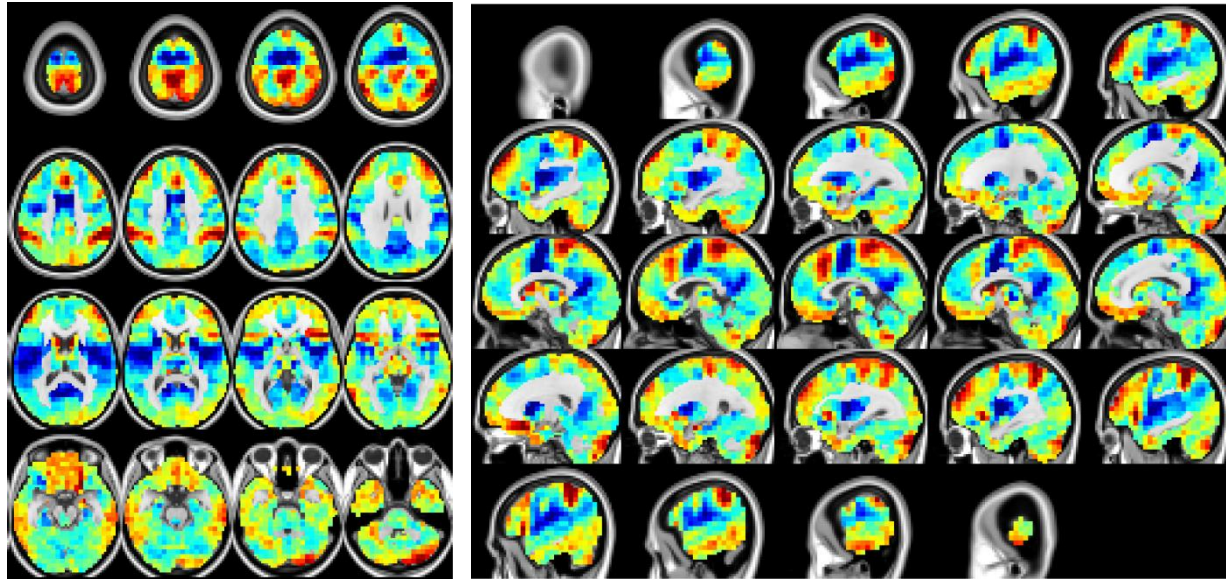
Thread 5
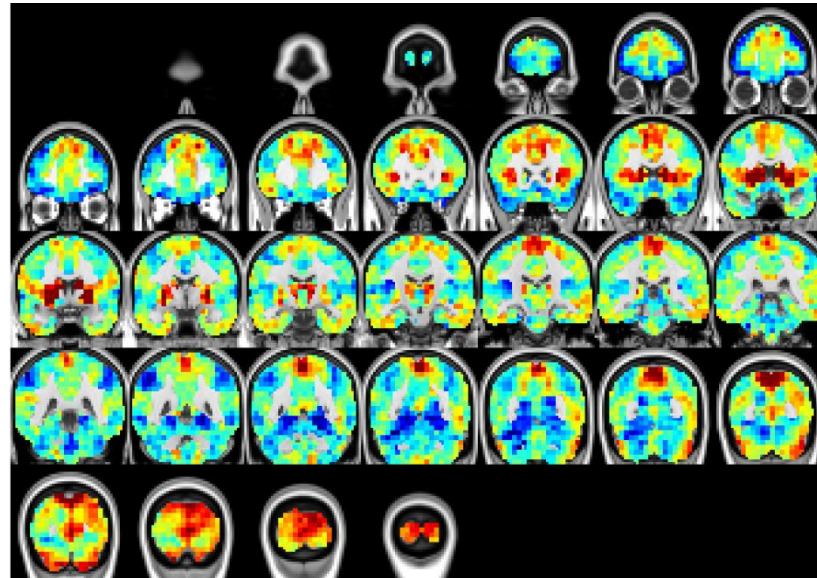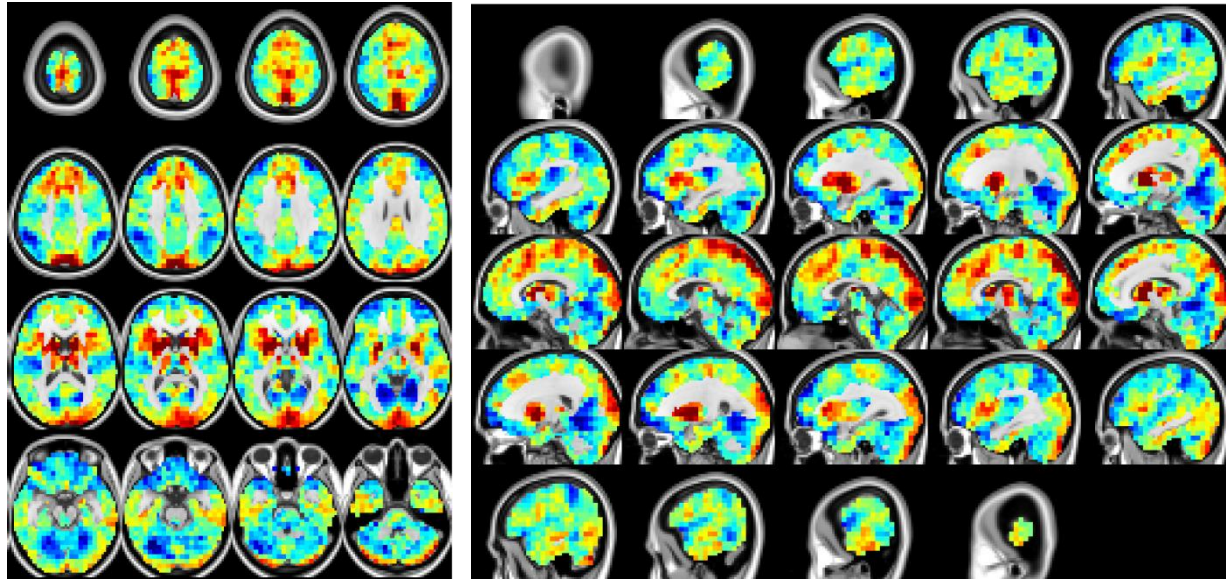
Thread 6
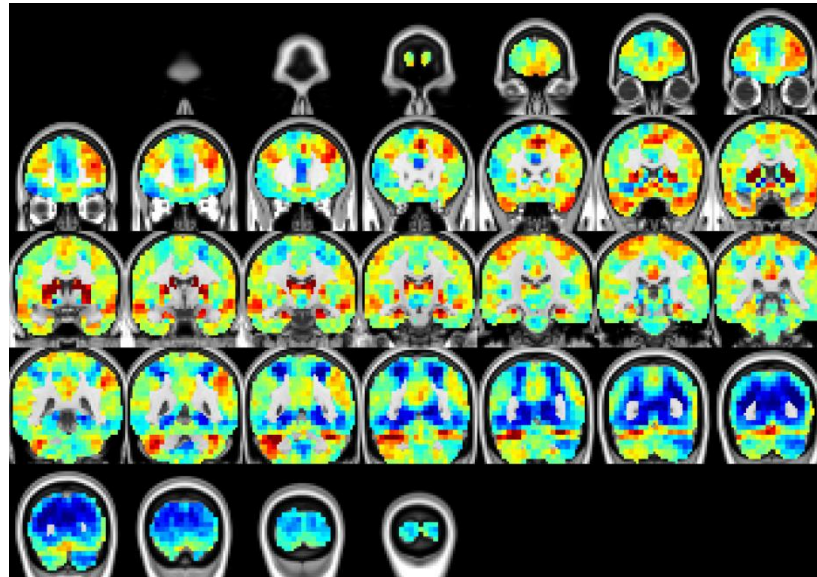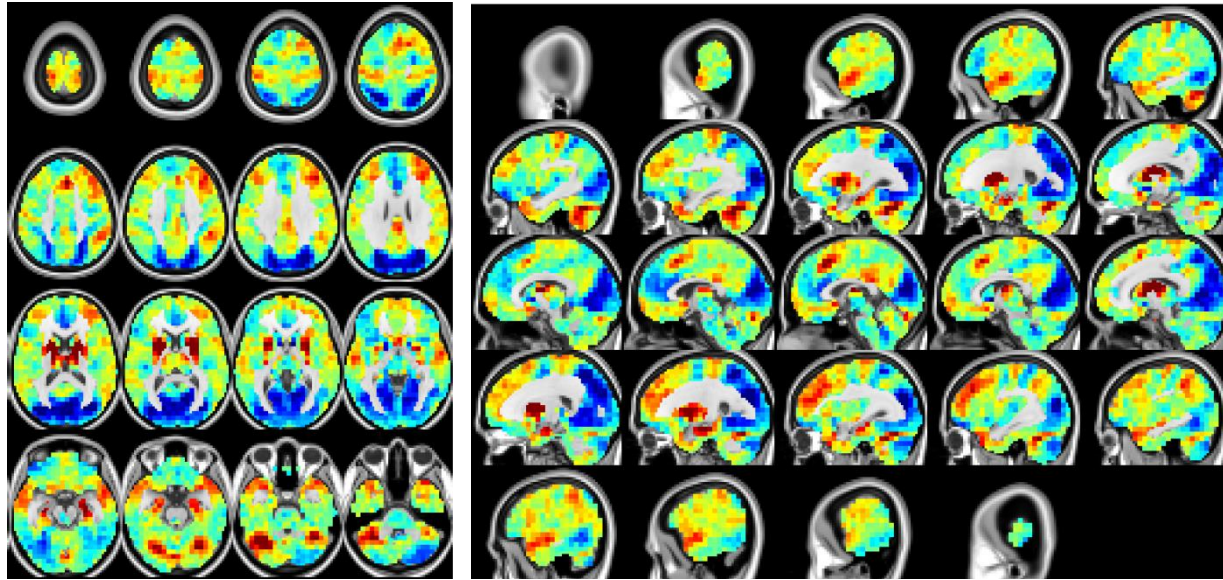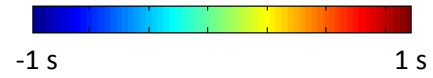
Thread 7
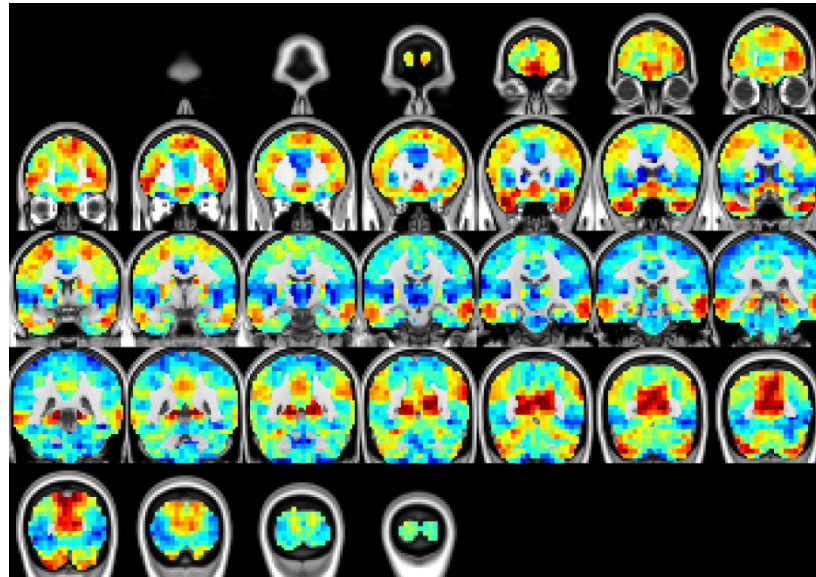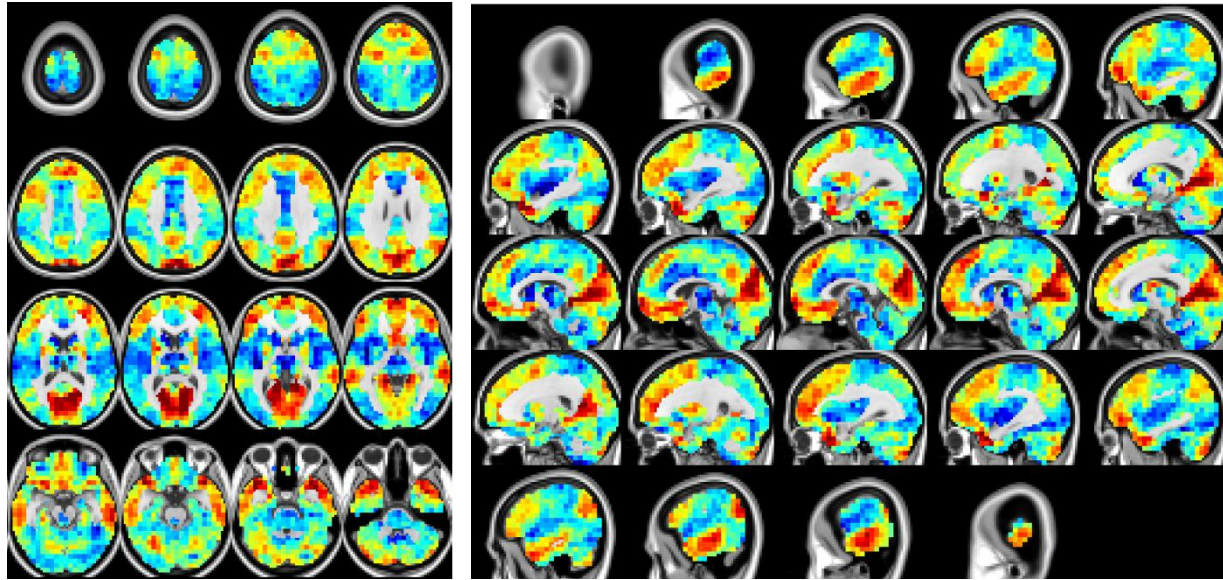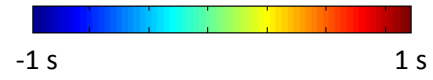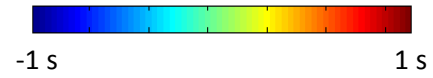
Thread 8
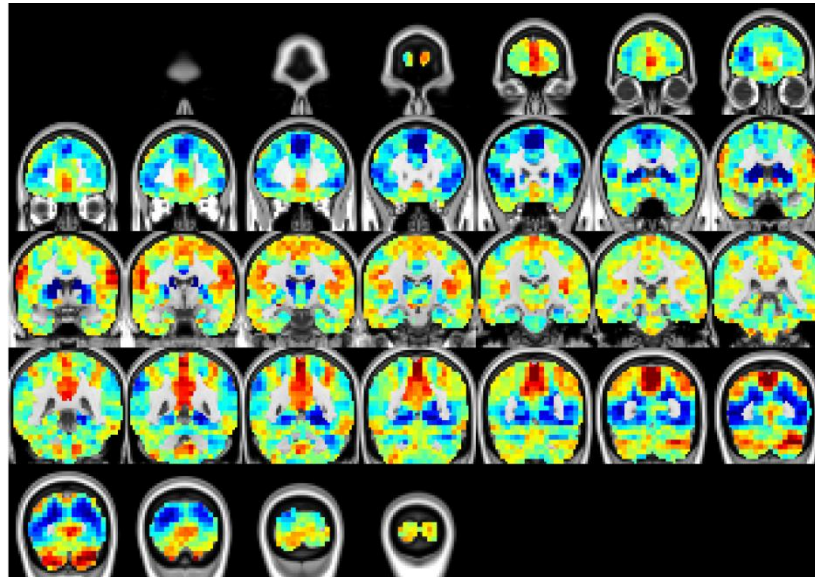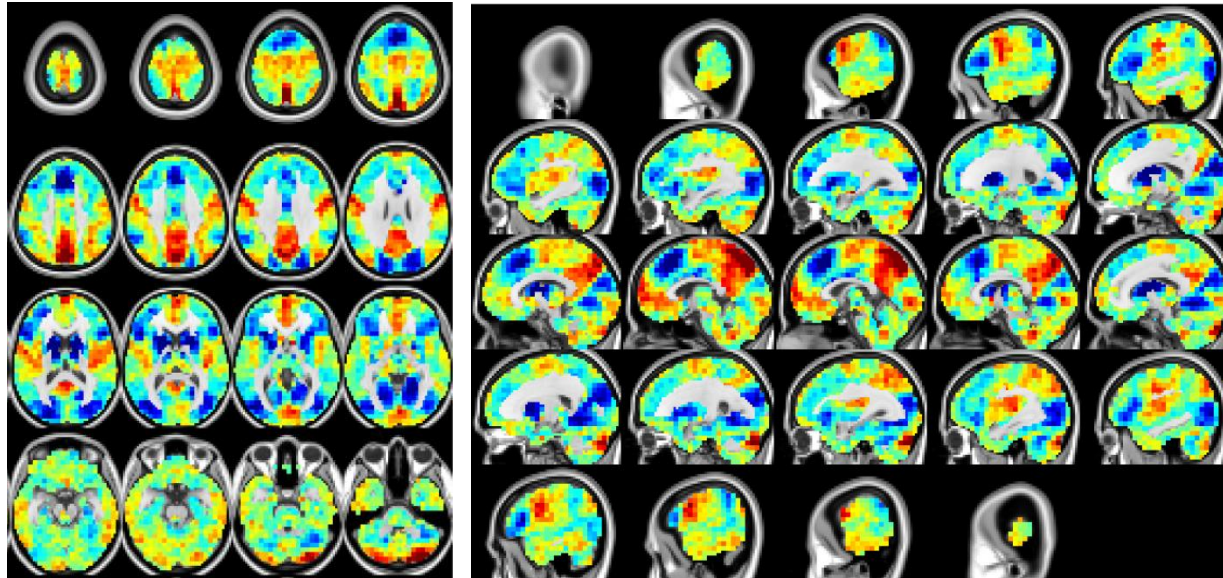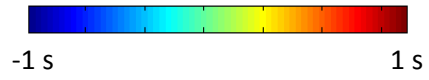
Anterior insula lag map

Posterior insula lag map

-1 s   1 s
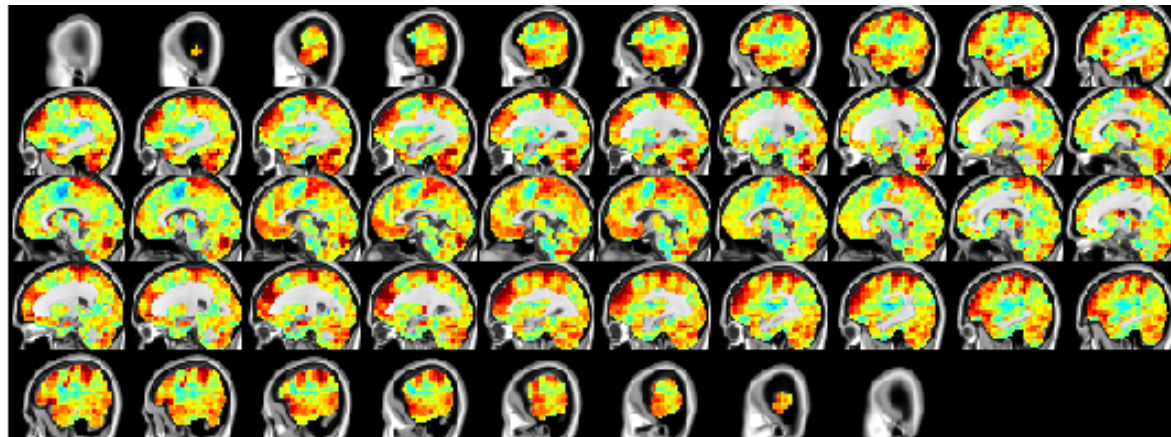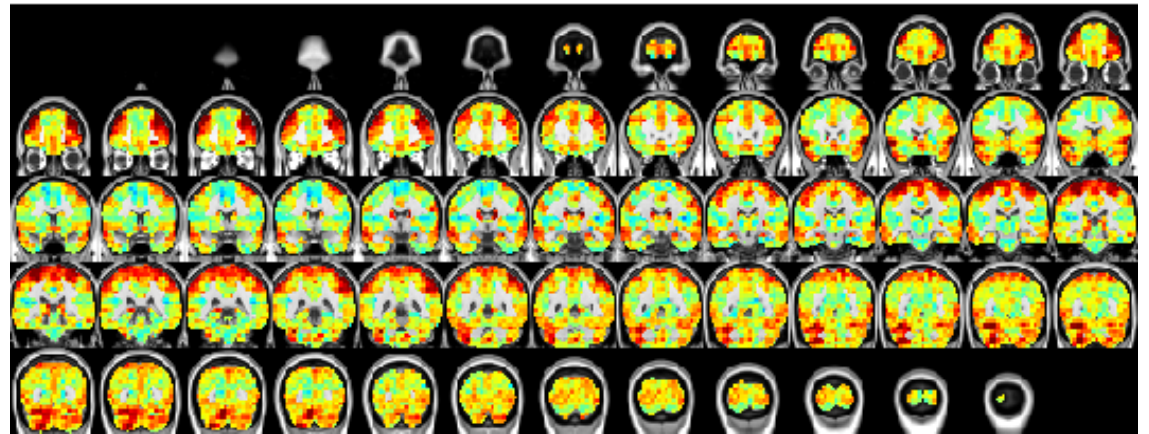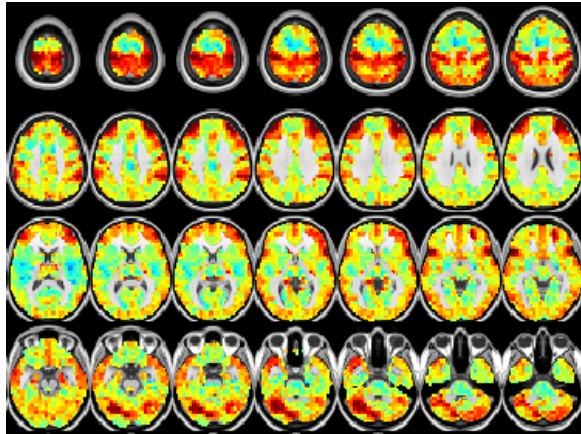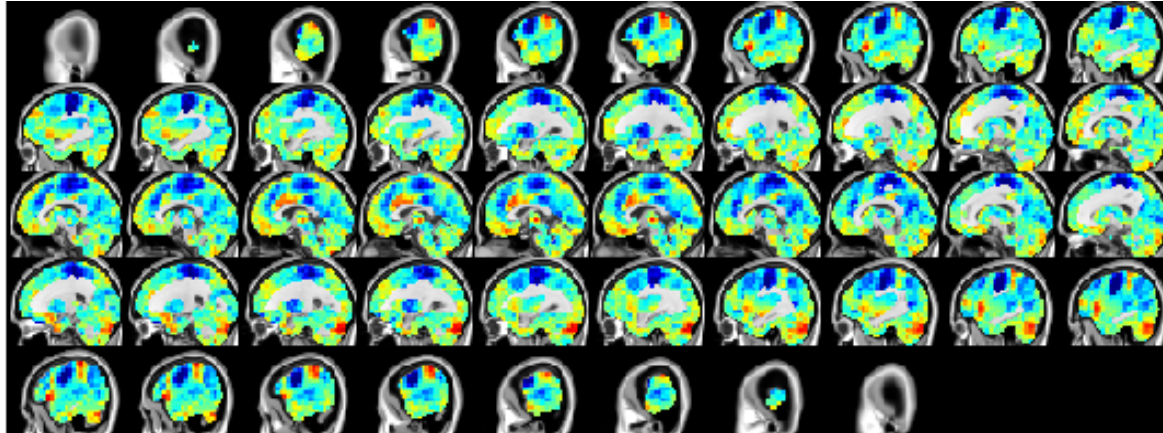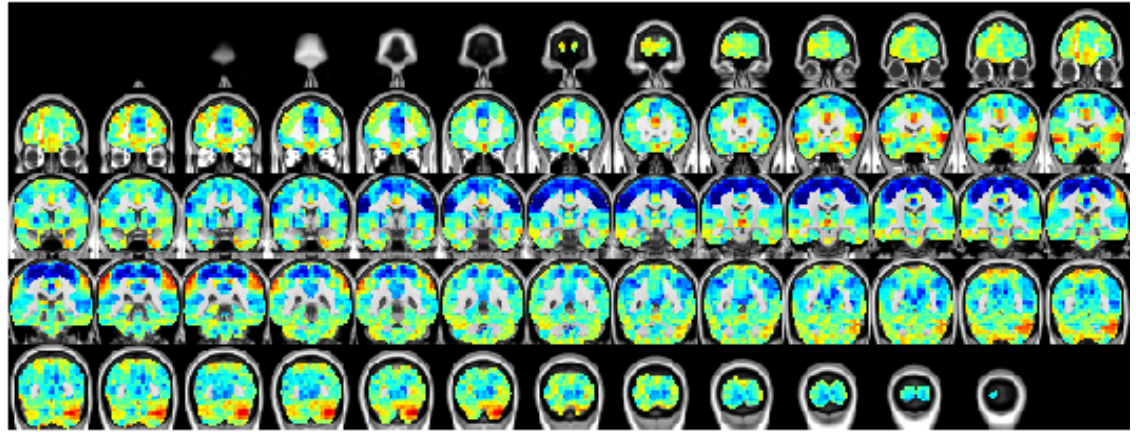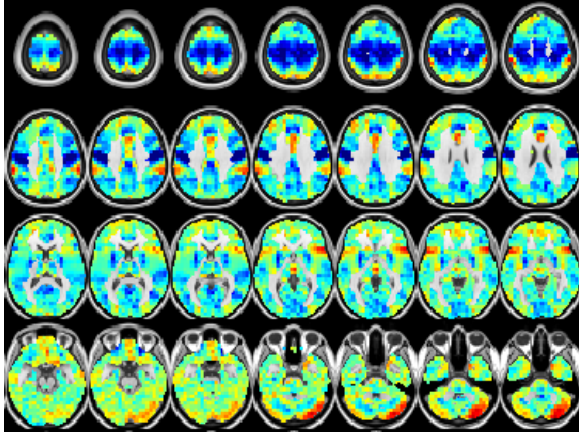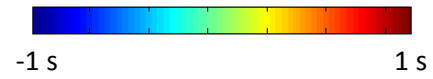
Caudate lag map
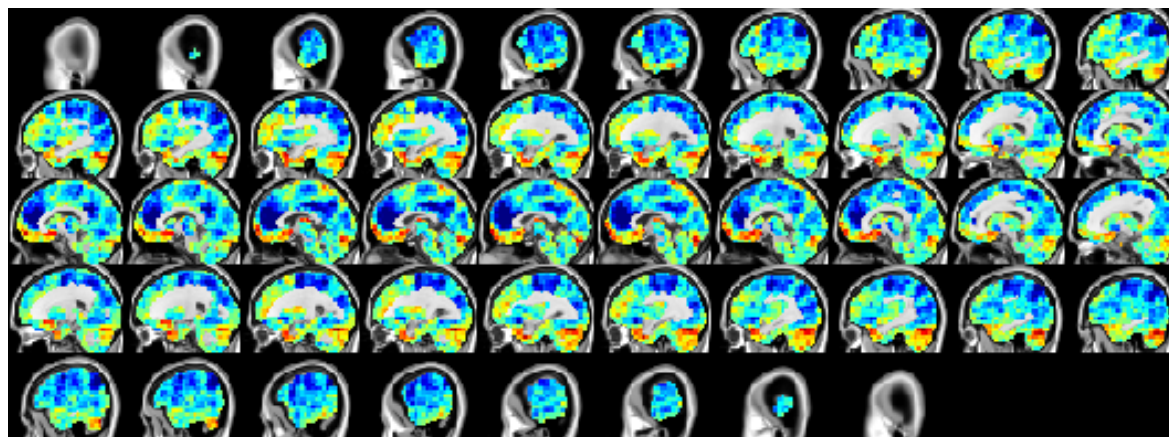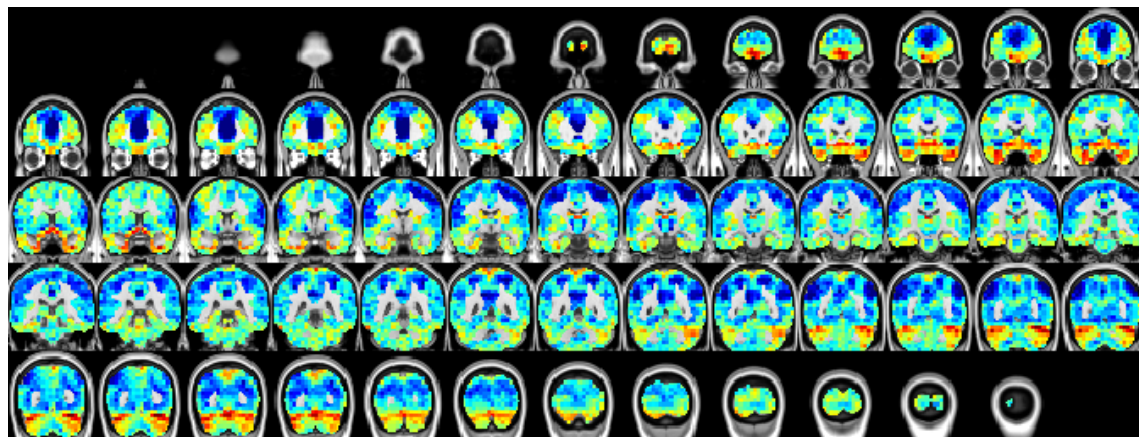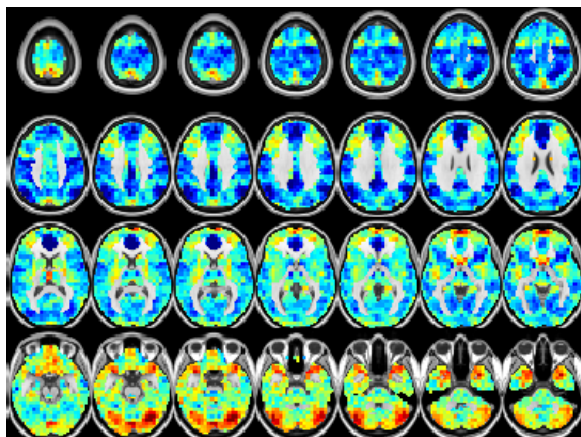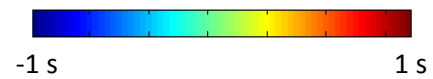
Putamen lag map
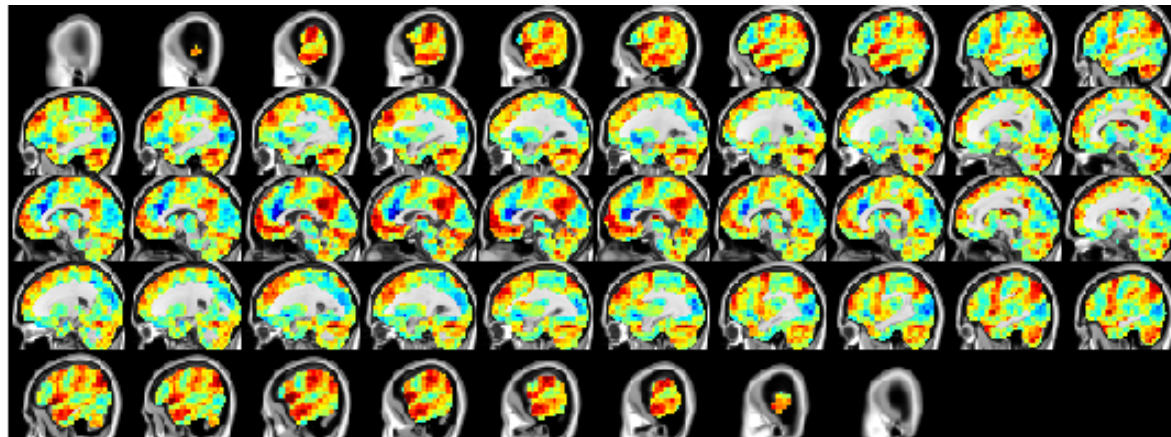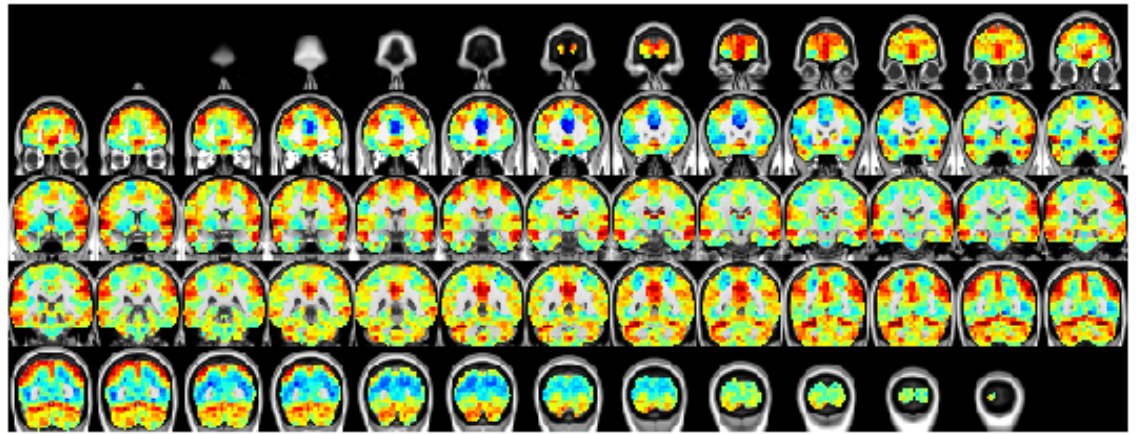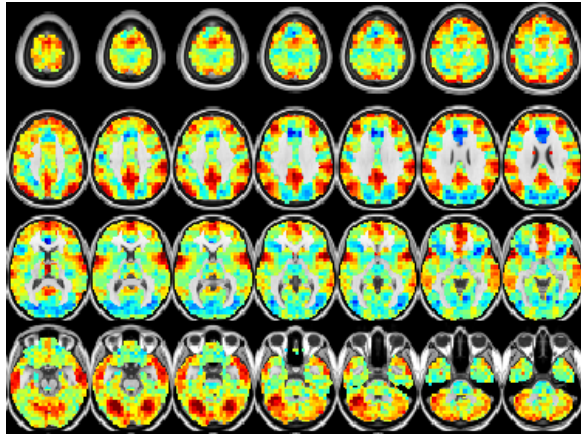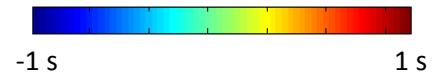
Entorhinal cortex lag map
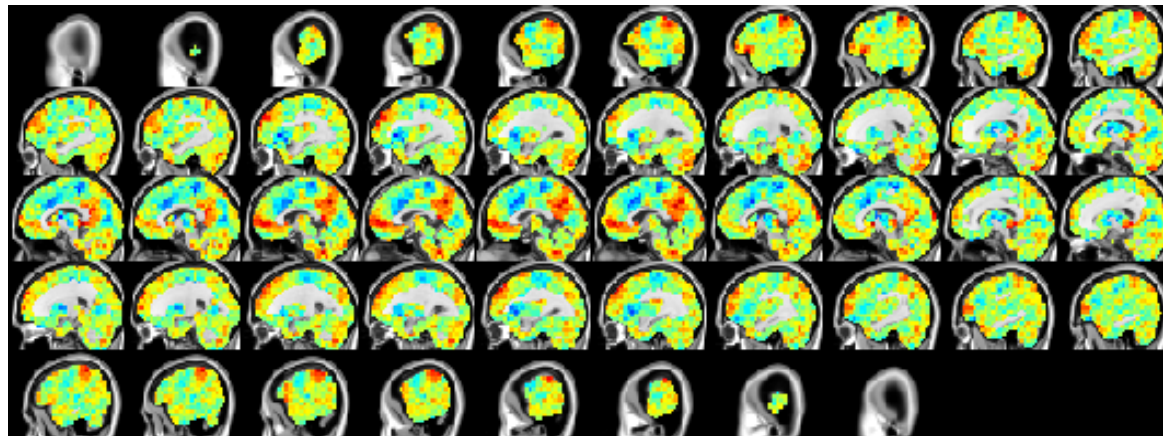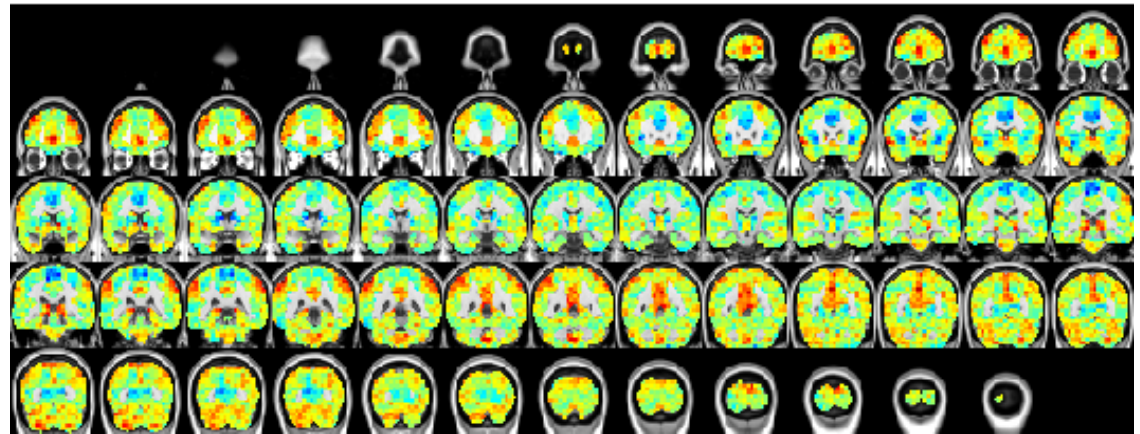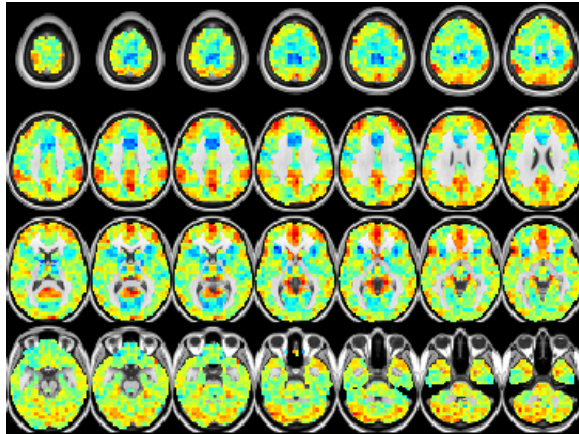
Medial prefrontal cortex lag map
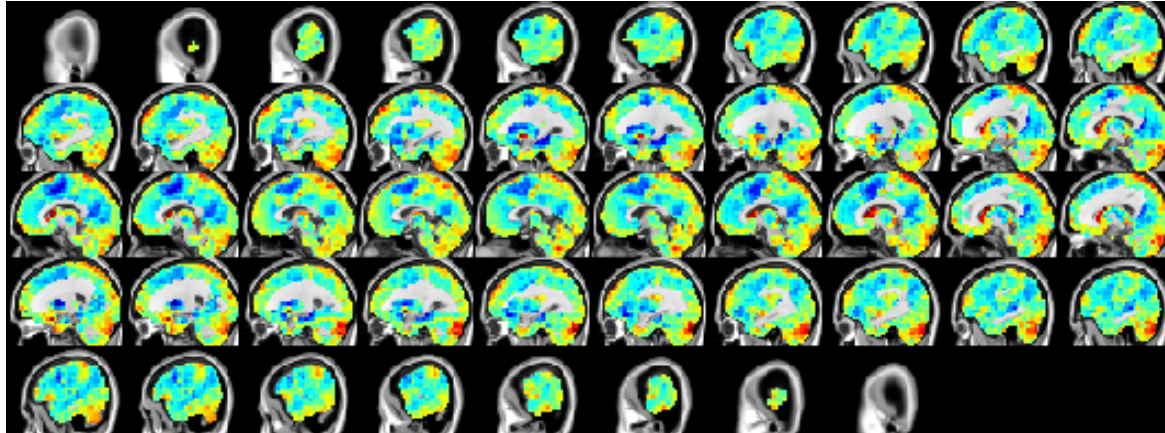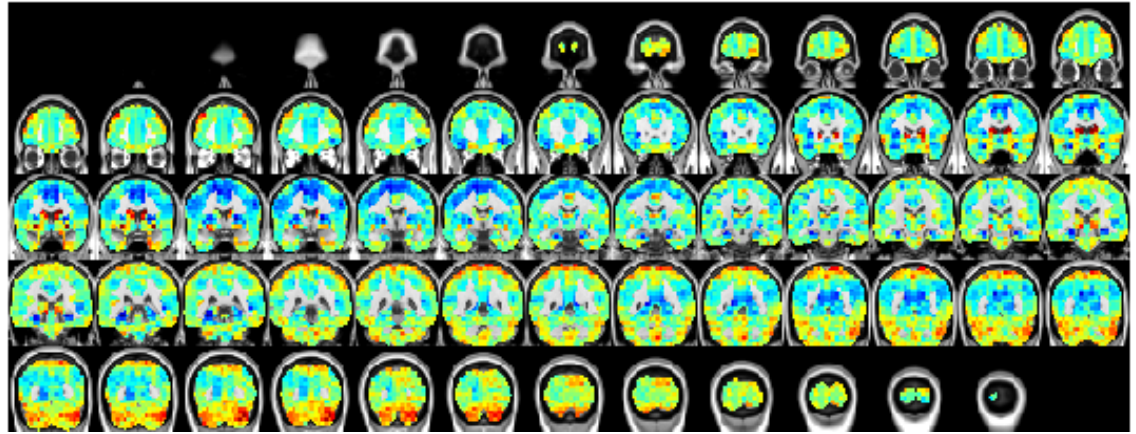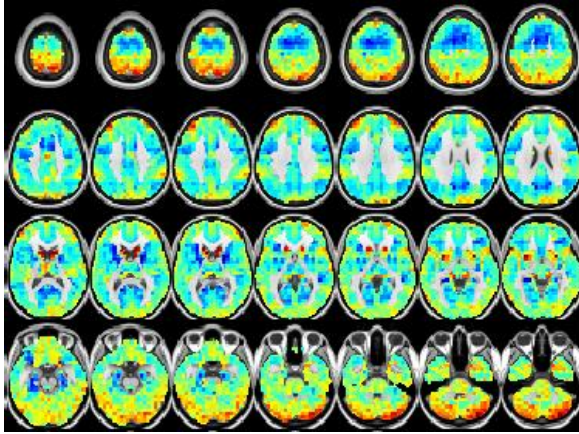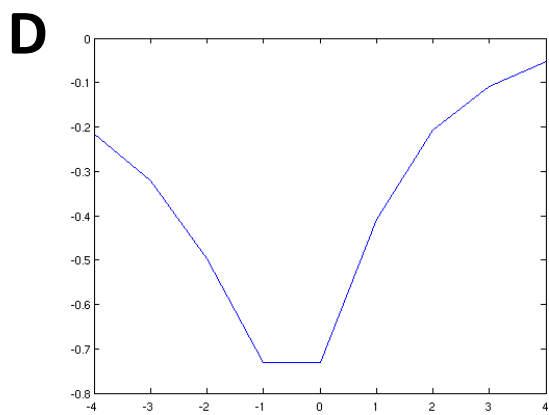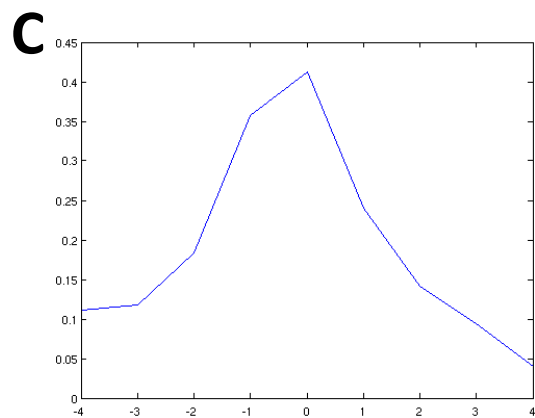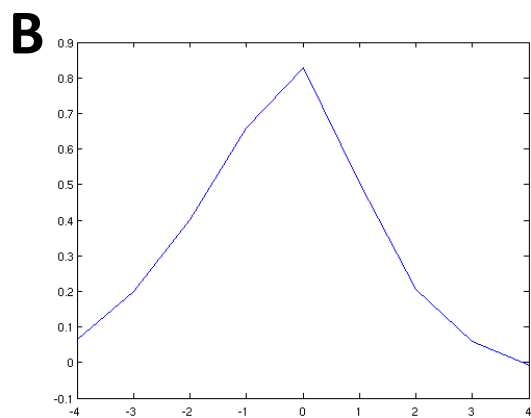
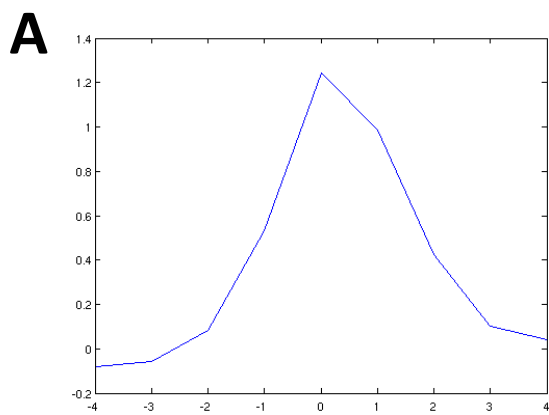# Posterior cingulate precuneus cortex lag map
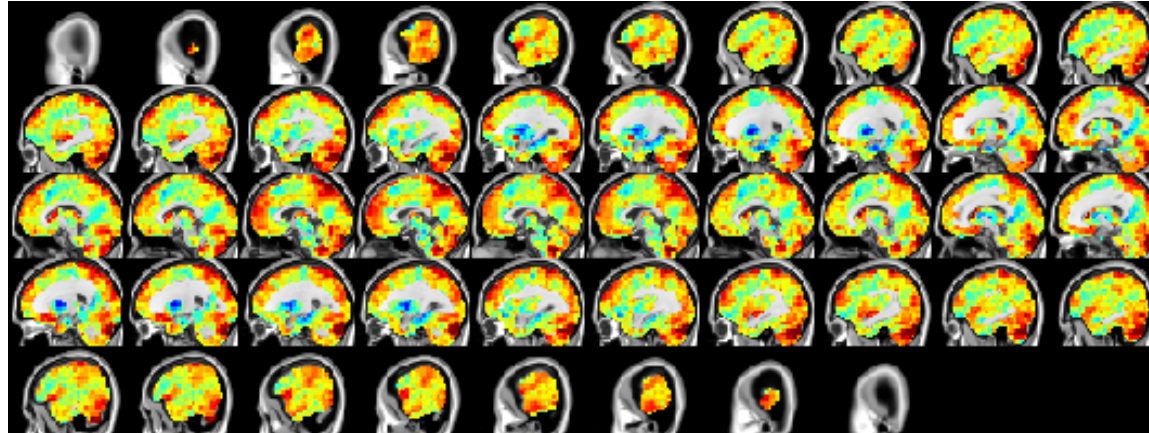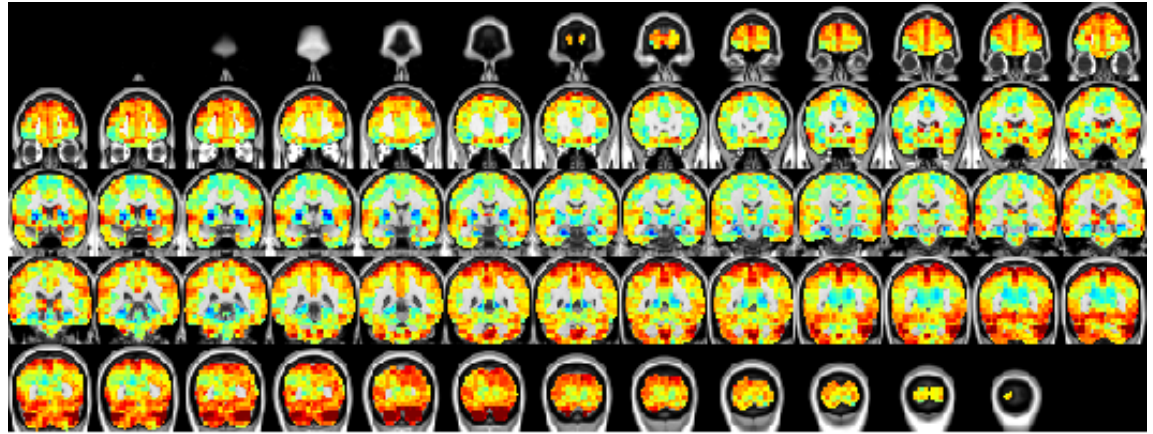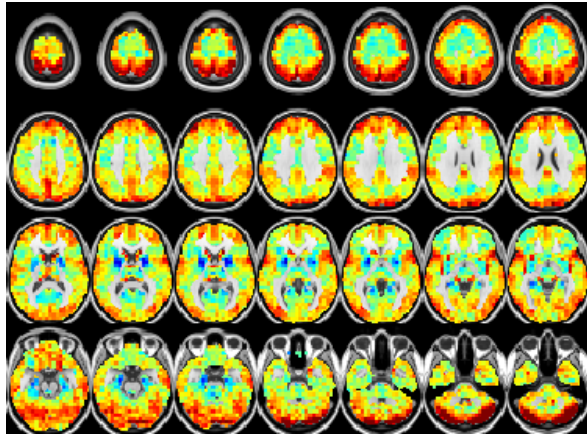
# Premotor cortex lag map
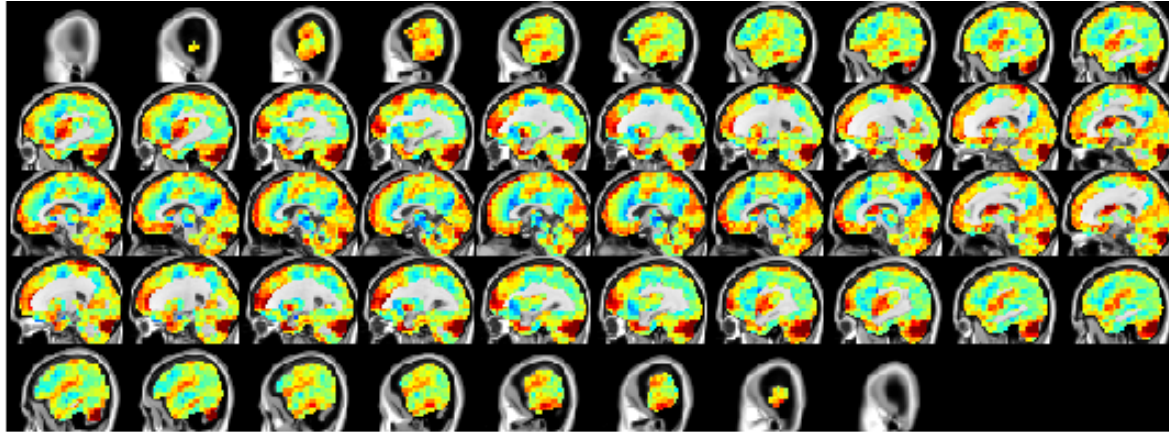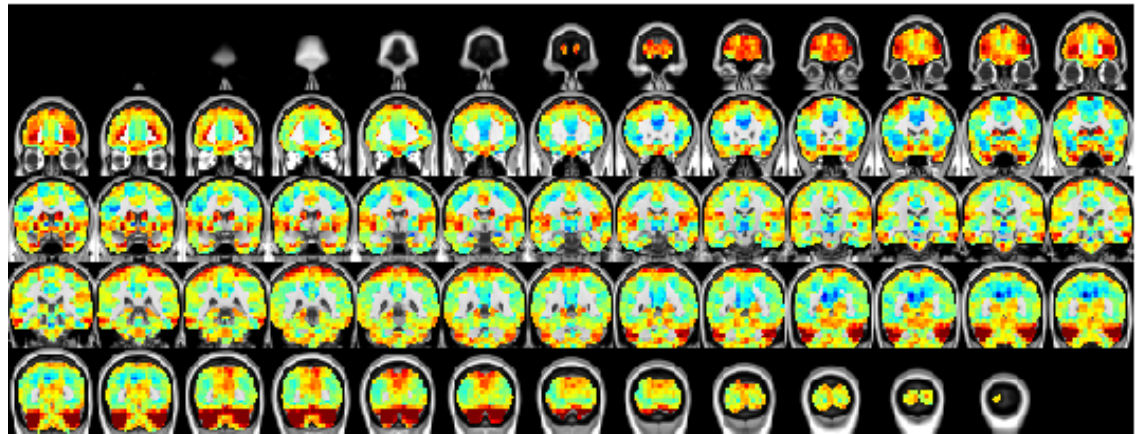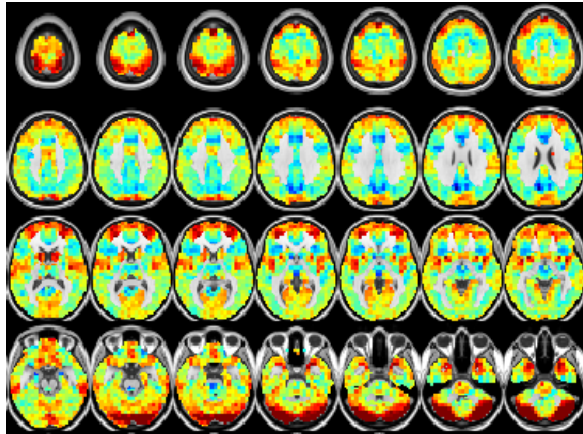
Supplementary motor area lag map

Parietal cortex lag map

Lagged covariance curve between
nodes 1 and 3
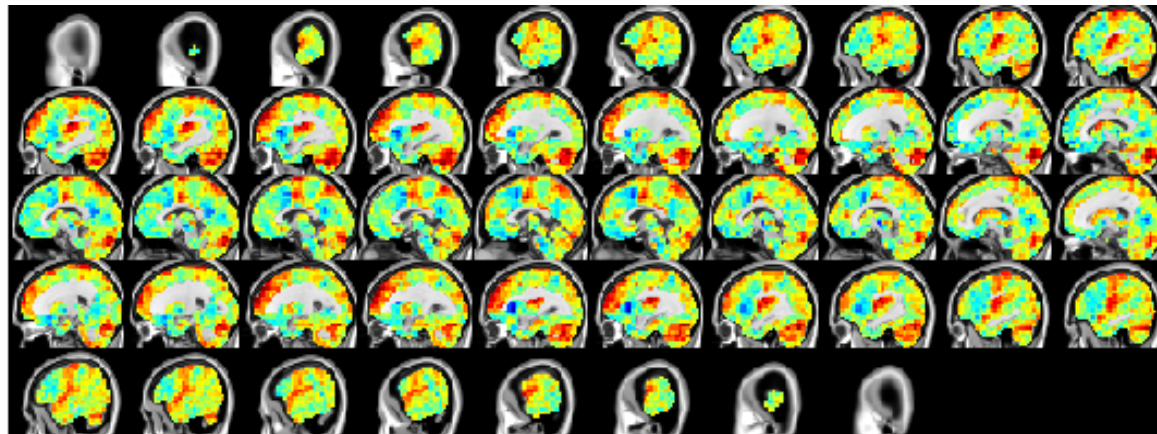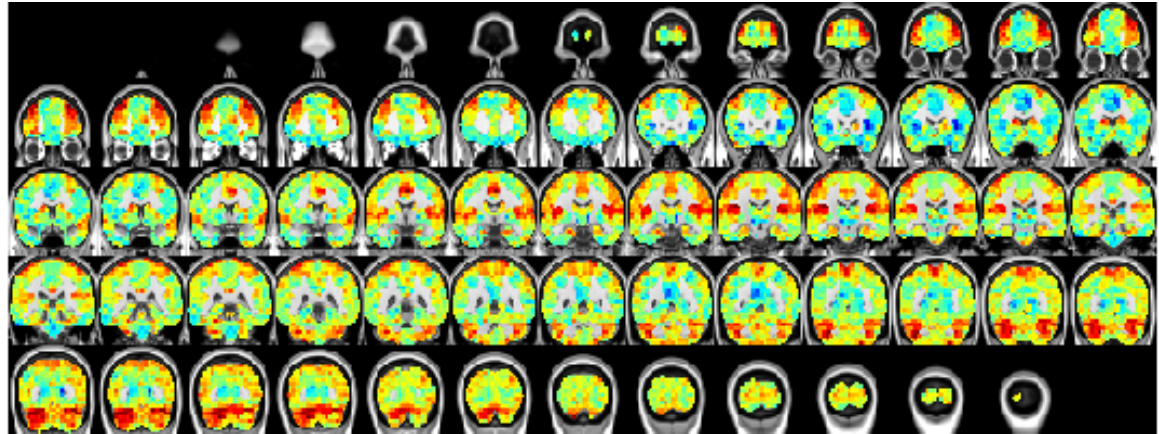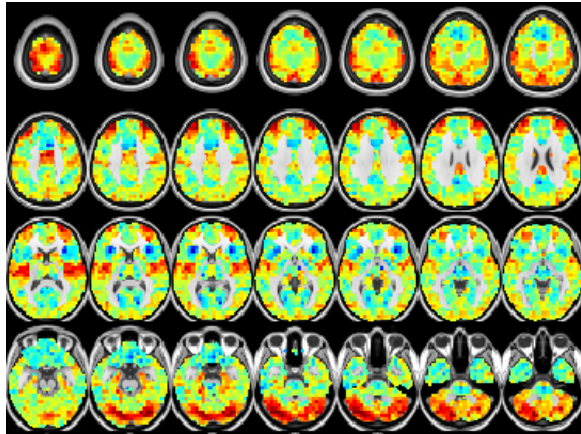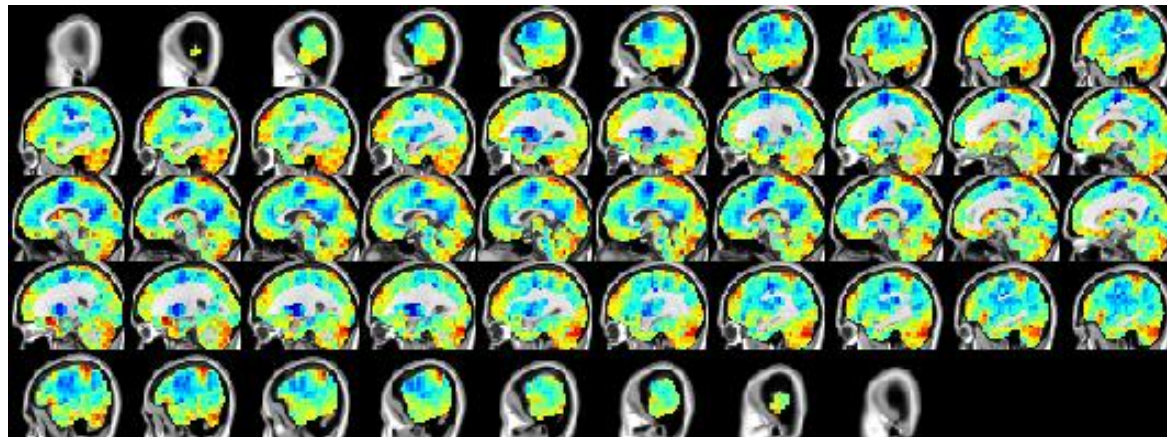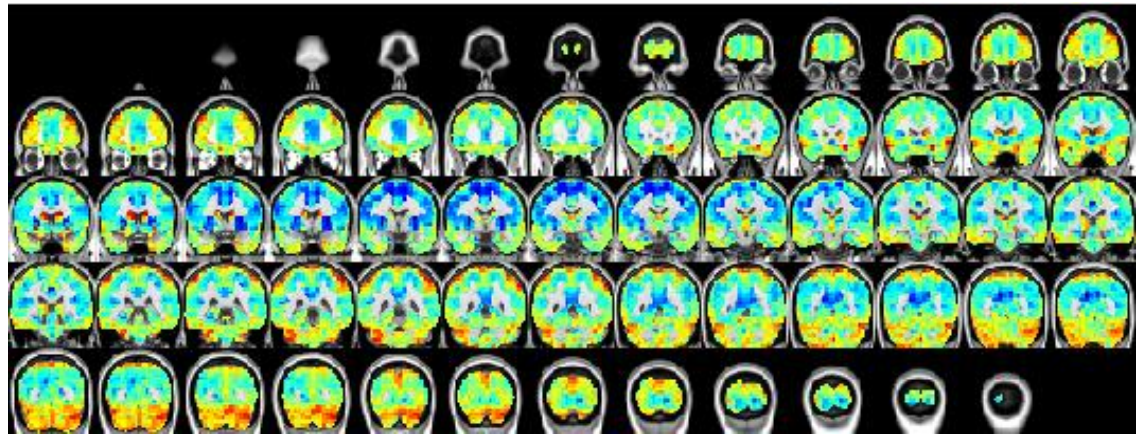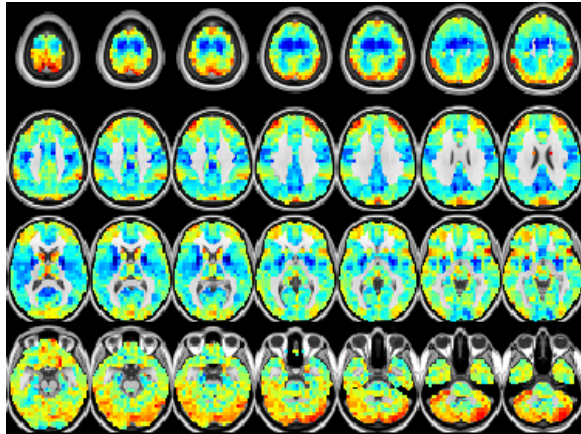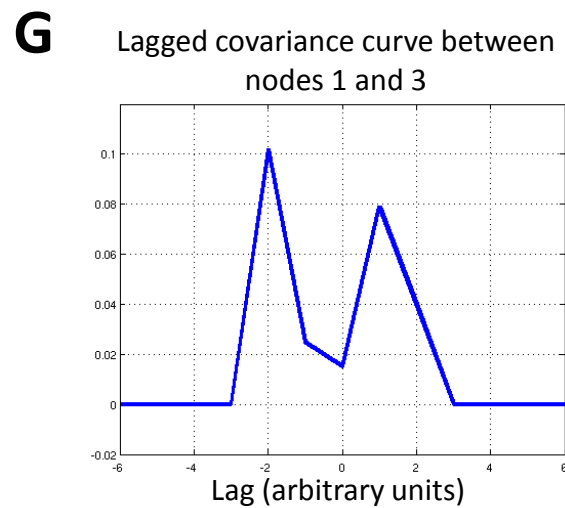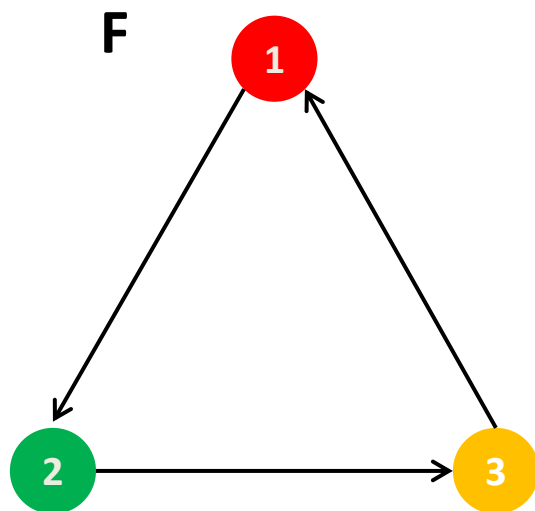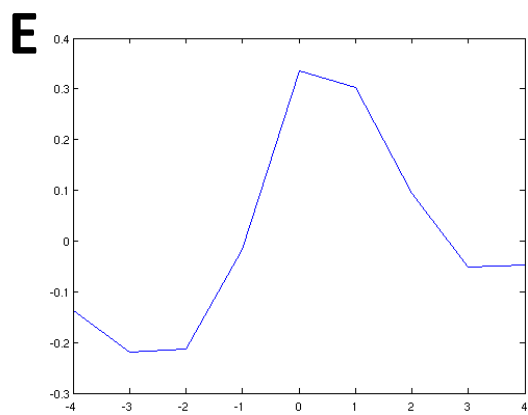
Averaging cross-covariance curves vs. averaging TD matrices does not materially affect the results– demonstrated using 17× 17 case

r = 0.98

Lag values when averaging over individual TD matrices

Lag values when averaging cross-covariance curves
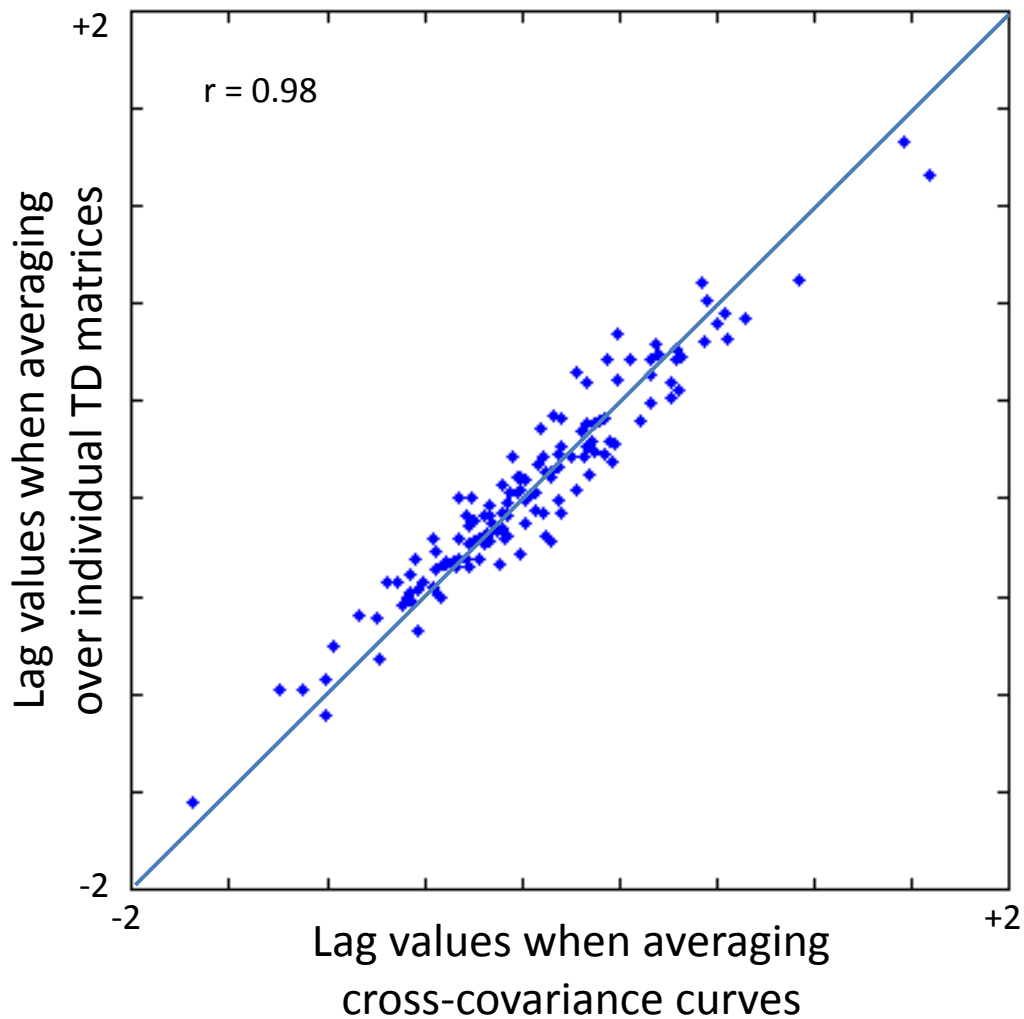
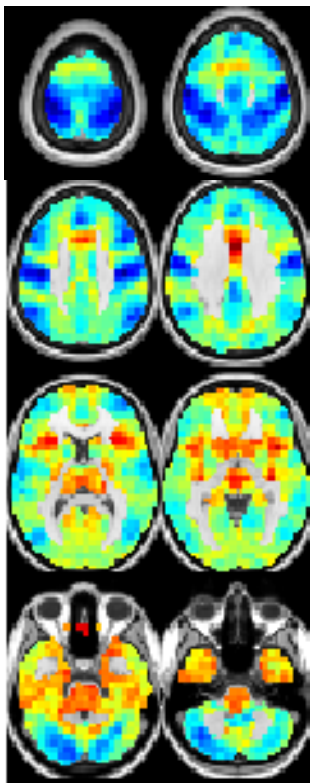**A** Convergence in the 17× 17 case
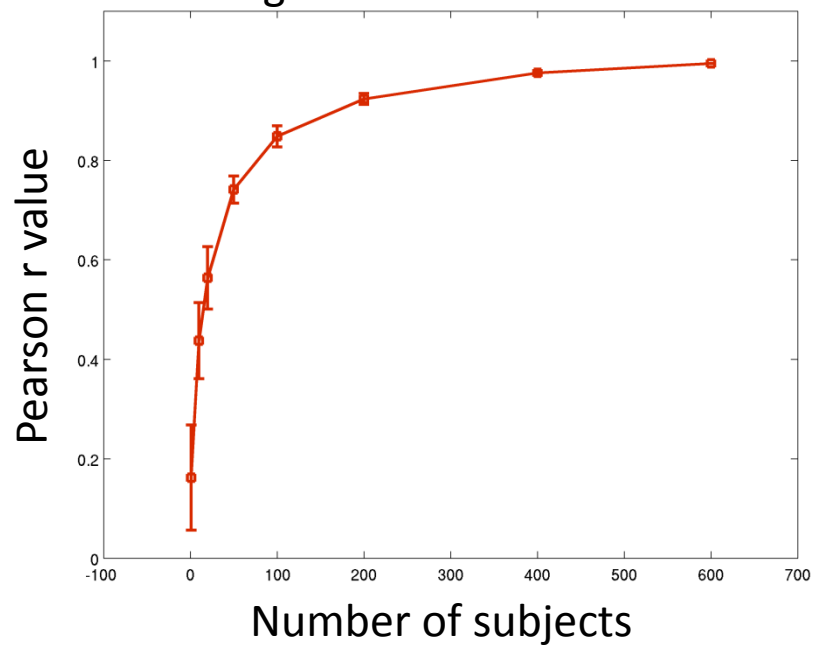
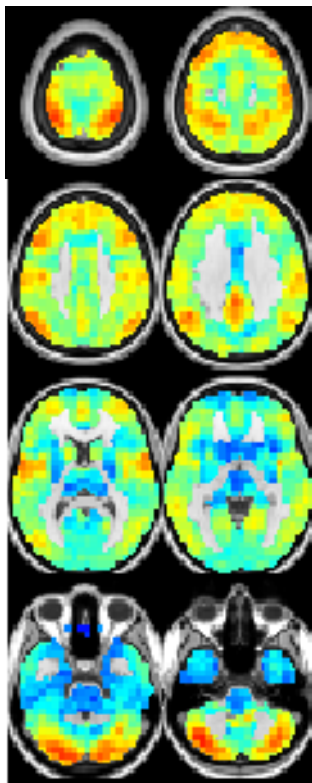**B** Convergence in the 6526 × 330 case
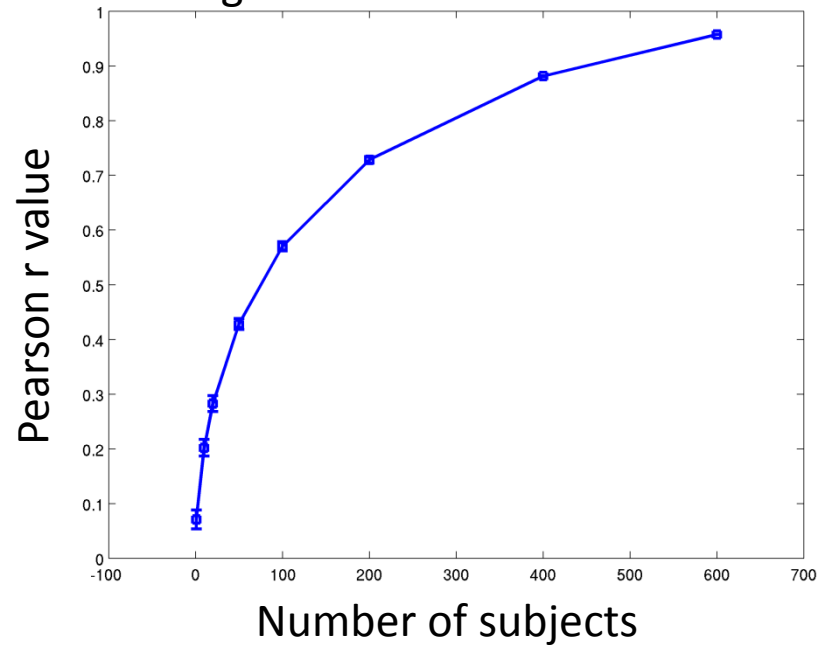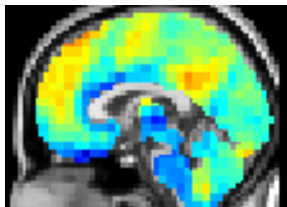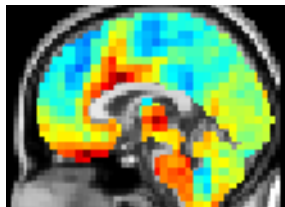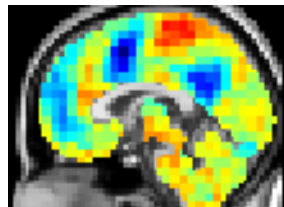
KPC 1   KPC 2   KPC 3   KPC 4
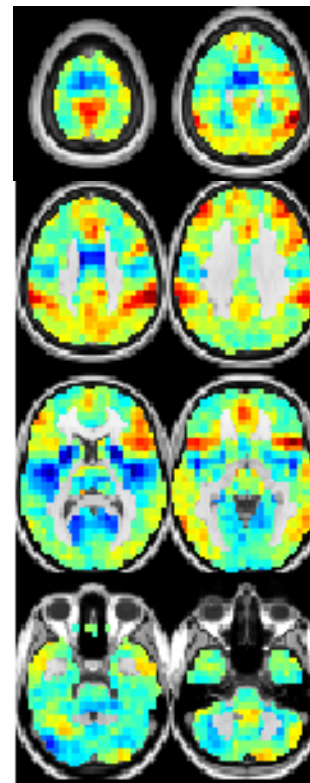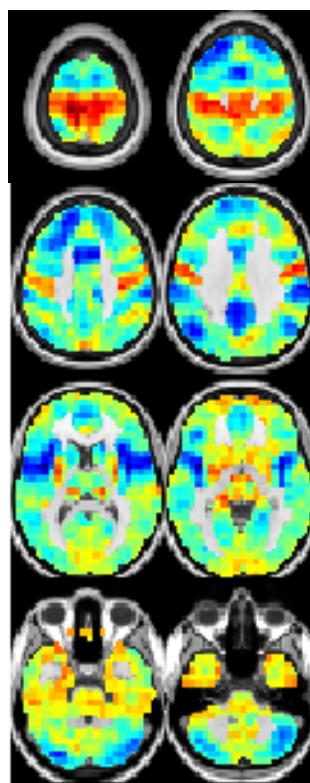
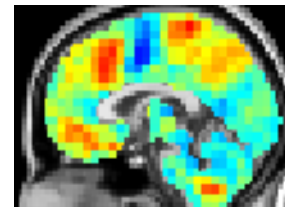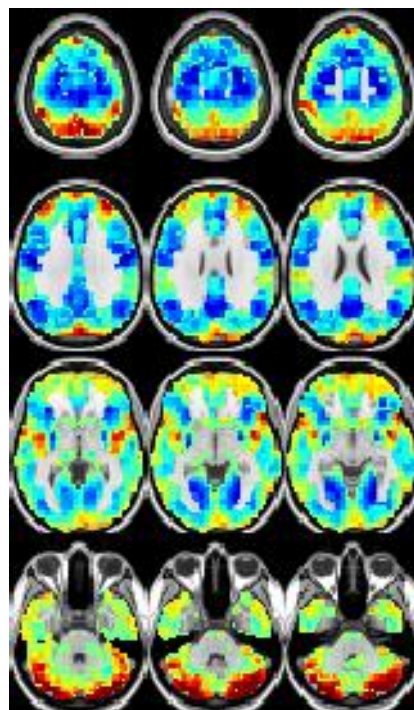**A**   Lag projection   **B**   Weighted sum of lag threads