Functional Specialization and Flexibility in Human Association Cortex

Supplemental Material

This supplemental material is divided into *Supplemental Methods* and *Supplemental Results* to complement the Methods and Results sections in the main text, respectively.

Supplemental Methods

Author-topic hierarchical Bayesian model

The details of the model can be found elsewhere (Rosen-Zvi et al., 2010). For the sake of completeness, further details about the model pertaining to this work are described in this section. In our application, the model parameters are the probability that a task will recruit a component (θ, i.e., Pr(component | task)) and the probability that a recruited component will activate a voxel (β, i.e., Pr(voxel | component)). θ and β are matrices, where each row is a categorical distribution summing to one.

The model assumed symmetric Dirichlet(α) and Dirichlet(η) priors on the θ and β respectively (Figure S1). Let T_e denote the set of tasks (consisting of one or more BrainMap-defined task categories) employed in the e-th experimental contrast[1]; e took on values from 1 to 10449. The e-th experiment resulted in a binary activation image obtained from preprocessing the BrainMap data (see BrainMap subsection in the Methods section). The set of voxels with values of 1 corresponded to an expanded set of activation foci associated with the experiment. The expanded set of activation foci were assumed to be independently and identically generated[2] (Rosen-Zvi et al, 2010). More specifically, the location v_{ef} of the f-th activation focus was assumed to be generated as follows:

1. Randomly generate a task T_{ef} uniformly from the set of tasks T_e.

[1] A small percentage of experiments in BrainMap utilized tasks from more than one task category.

[2] Note that the locations of the activation foci are independent conditioned on knowing θ and β. However, the locations of the activation foci are not independent if θ and β are unknown.

2. Randomly generate a component C_{ef} using the distribution specified in the T_{ef}-th row of the θ matrix.
3. Randomly generate the voxel location v_{ef} of the activation focus using the distribution specified in the C_{ef}-th row of the β matrix.

The above description specified the model exactly (Rosen-Zvi et al., 2010).

Estimating the author-topic model

Given the 10,449 BrainMap experiments with associated activation coordinates and task categories, Pr(component | task) θ and Pr(voxel | component) β were estimated assuming the following fixed parameters: the number of cognitive components N_C , and the hyperparameters α and η.

The choice of N_C is discussed in the main text. Following the original author-topic paper (Rosen-Zvi et al., 2010), the hyperparameters α and η were set to $50/N_C$ and 0.01 respectively. Our experiments (not shown) suggested that the estimates were robust to the exact choice of α and η.

The original author-topic paper (Rosen-Zvi et al., 2010) utilized the Gibbs sampling algorithm to estimate Pr(topic | author) and Pr(words | topic). These are equivalent to Pr(component | task) θ and Pr(voxel | component) β here. However, the Gibbs sampling algorithm is too computationally inefficient for this particular dataset: the algorithm did not converge after one week of computation time. This inspired us to derive a much faster expectation-maximization (EM) algorithm (Dempster et al., 1977).

Given initial estimates of Pr(component | task) θ^0 and Pr(voxel | component) β^0, the E-step and M-step were iterated until convergence. In particular, let θ^i and β^i be the estimates of Pr(component | task) and Pr(voxel | component) after the i-th iteration of the EM algorithm. The $(i+1)$-th iteration of the E-step involved computing the posterior probability that the f-th activation focus (located at voxel v_{ef} from the binary activation image of the e-th experiment) was generated by task t and component c:

$$q_{ef}^{i+1}(t,c) \propto \frac{\delta(t \in T_e)}{|T_e|} \theta_{t,c}^{i} \beta_{c,v_{ef}}^{i} ,$$

where $\delta(t \in T_e)$ was equal to one if task t belongs to the set of tasks T_e employed in the e-th experiment and was equal to zero otherwise. $|T_e|$ indicated the size of the task set T_e. $\theta^i_{t,c}$ corresponded to the t-th row and c-th column of θ^i, while $\beta^i_{c,v_{ef}}$ corresponded to the c-th row and v_{ef}-th column of β^i. The posterior probability $q^{i+1}_{ef}(t,c)$ was then used to estimate θ^{i+1} and β^{i+1} in the $(i+1)$-th iteration of the M-step:

$$\theta^{i+1}(t,c) \propto \sum_{e=1}^{E}\sum_{f=1}^{F_e}\left(q^{i+1}_{ef}(t,c) + \alpha\right)$$

$$\beta^{i+1}(c,l) \propto \sum_{e=1}^{E}\sum_{f=1}^{F_e}\left[\left(\sum_{t\in T_e} q^{i+1}_{ef}(t,c)\right) w^l_{ef} + \eta\right] \quad ,$$

where l indexed voxel locations within a MNI152 brain mask. The hyperparameters α and η acted as pseudo-observations, so that the estimates θ and β were non-zero for all tasks, components and locations within the mask. w^l_{ef} was a positive constant if activated voxel v_{ef} corresponded to location l, and zero otherwise. The exact value of w^l_{ef} was not important; what mattered was the ratio between w^l_{ef} and η. In other words, as long as w^l_{ef} and η were scaled by the same amount, the estimates would always be the same. Here w^l_{ef} was set to five.

Although the EM algorithm is much faster, we found that the Gibbs sampling algorithm provided qualitatively better results, possibly because Gibbs sampling was able to explore a larger portion of the parameter space. Consequently, the resultant estimation procedure was as follows: (1) Gibbs sampling was performed for 100 iterations (a few hours) and (2) the EM algorithm was run with the Gibbs sampling output as initialization (a few hours). The estimation procedure was repeated with 100 random initializations resulting in 100 estimates.

A final estimate was obtained by selecting the solution closest to the remaining 99 estimates. Briefly, for each pair of estimates, we reordered the components (using the Hungarian matching algorithm) to maximize the correlations of the Pr(voxel | component) between corresponding pairs of components. After obtaining the optimal match, the pairwise correlations were averaged across all components, resulting in an average correlation value indicating the quality of match between the pair of estimates. The estimate resulting in the highest average

correlation values with the remaining 99 estimates was taken as the final estimate. We obtained the same final estimate using KL-divergence instead of correlation.

Exhaustive search of nested ontology

An exhaustive search was performed to quantify the scenario that two components of the (N+1)-component estimate were subdivisions of a component of the N-component estimate (while the remaining N-1 components remained similar across both estimates). This quantification is based on the following idea: if a component of the N-component estimate divides into i-th and j-th subcomponents of the (N+1)-component, then the combination of the two subcomponents should be similar to the original component. To quantify the presence of this phenomenon, the Pr(voxel | component) of the i-th and j-th components were averaged[3] into a single Pr(voxel | component). The resulting N components of the (N+1)-component estimate were matched to the N-component estimate by reordering the components (using the Hungarian matching algorithm) to maximize the correlation of the Pr(voxel | component) between corresponding pairs of components. After obtaining the optimal correspondence, the pairwise correlations were averaged across all pairs of components, resulting in an average correlation value indicating the quality of the split (with higher correlation values indicating a better split). By performing an exhaustive search over all values of i and j, we found the component of the N-component estimate whose split best approximates the (N+1)-component estimate. This procedure was independently repeated using Pr(component | task). The estimated splits were visually inspected. If different (i, j) pairs were found via exhaustive search of Pr(voxel | component) and Pr(component | task), visual inspection was used to resolve the differences.

An alternative measure of flexibility: unnormalized entropy

To ensure our analyses are robust to the exact definition of flexibility, unnormalized entropy (defined as $-\sum_i Pr(voxel \mid component\ i) * log\ Pr(voxel \mid component\ i)$) was also considered as an alternative measure. The Shannon entropy measure (defined as $\sum_i -Pr(component\ i \mid voxel) * log\ Pr(component\ i \mid voxel)$) was deemed unsuitable due to the following toy example. Consider a voxel with Pr(voxel | component) equal to 1e-6 for all

[3] A nice property here is that the averaged Pr(voxel | component) is still a valid probability distribution.

components and a second voxel with Pr(voxel | component) equal to 1e-5 for all components. Then Pr(component | voxel) will be equal to $1/N_C$ for both voxels for both components. Therefore the entropy measure will consider both voxels as equally flexible, while the unnormalized entropy measure will consider the second voxel to be more flexible.

An alternative model

Here we consider an alternative generative model with N cognitive components. For simplicity, each experiment was assumed to only utilize one task. The activations of an experiment utilizing task T were assumed to be generated using a two-step procedure. In the first step, the set of components recruited during task T was randomly generated via the probability distributions Pr(component | task). Here, there are N probability distributions per task, where Pr(k-th component | task T) is a number between zero and one, specifying the probability that the k-th component is going to be active or not during task T. This contrasts with the author-topic model, where there is one probability distribution per task Pr(component | task), which consists of N numbers (between zero and one) that sum to one.

In the second step, each brain voxel was then determined to be activated if any of the recruited components (from the first step) activated the voxel according to Pr(voxel | component). Here, each component has an associated Pr(voxel | component) for each voxel. Pr(voxel | component) is a number between zero and one for each voxel (and for each component). This contrasts with the author-topic model, where Pr(voxel | component) is a number between zero and one that sums to one over all voxels (for each component). For example, if components C1 and C2 were recruited (according to the first step) and Pr(voxel 1 | C1) = 0.8 and Pr(voxel 1 | C2) = 0.3, then the probability that voxel 1 was not activated is (1 – 0.8)*(1 – 0.3).

It is unclear whether this alternative model is biologically more plausible than the author-topic model, although the probabilities may be interpreted without the qualification “for an activation focus”, as in the case for the author-topic model.

We had previously experimented with a variation of this model, but the resulting algorithm was many times slower than the author-topic model. The reason was that the author-topic model generates brain activation by iterating over activation foci, and therefore the resulting estimation algorithm iterated over activation foci. The current inference algorithm

(using Gibbs sampling and EM) took a few hours per random initialization. By contrast, the alternative model generates brain activation by iterating over all brain voxels in MNI152 space, and therefore the resulting estimation algorithm iterated over all brain voxels in MNI152 space. Given that there were many times more brain voxels than the number of activation foci, the resulting algorithm for the alternative model was orders of magnitudes slower. When the alternative model was initialized with the author-topic estimate, the resulting estimate was qualitatively similar. Because of the computational complexity, the alternative model has *not* been estimated with random initializations. Developing more efficient algorithms for estimating the alternative model will be left for future work.

Supplemental Results

Table S1: Examples of nested ontology of Pr(component | task)

n components	n+1 components	
Pr(^{10}C9 | task)	**Pr(^{11}C9 | task)**	**Pr(^{11}C10 | task)**
Theory of Mind: 0.57	Theory of Mind: 0.63	Face Mon/Discrim: 0.79
Episodic Recall: 0.52	Episodic Recall: 0.55	Olfactory Mon/Discrim: 0.45
Face Mon/Discrim: 0.52	Rest: 0.42	Passive Viewing: 0.39
Subj Emo Pict Discrim: 0.43	Deception: 0.34	Classical Conditioning: 0.36
Acupuncture: 0.40	Acupuncture: 0.34	Subj Emo Pict Discrimin: 0.31
Pr(^{11}C8 | task)	**Pr(^{12}C8 | task)**	**Pr(^{12}C9 | task)**
WCST: 0.50	Flanker: 0.49	WCST: 0.59
Counting/Calculation: 0.45	Deception: 0.45	Counting/Calculation: 0.45
n-back: 0.41	Go/No-Go: 0.39	n-back: 0.38
Sternberg: 0.40	Stroop: 0.31	Sternberg: 0.37
Flanker: 0.38	Simon: 0.27	Task Switching: 0.33
Pr(^{12}C4 | task)	**Pr(^{13}C4 | task)**	**Pr(^{13}C5 | task)**
Visual Pursuit/Tracking: 0.59	Naming (C): 0.55	Visual Pursuit/Tracking: 0.80
Action Observation: 0.58	Naming (O): 0.52	Action Observation: 0.59
Naming (C): 0.39	Face Mon/Discrim: 0.28	Mental Rotation: 0.35
Naming (O): 0.32	Passive Viewing: 0.18	Visual Distractor/Attn: 0.30
Mental Rotation: 0.32	Reading (C): 0.16	Saccades: 0.28
Pr(^{13}C11 | task)	**Pr(^{14}C11 | task)**	**Pr(^{14}C12 | task)**
Theory of Mind: 0.68	Rest: 0.49	Theory of Mind: 0.81
Episodic Recall: 0.40	Fixation: 0.37	Subj Emo Pict Discrim: 0.50
Rest: 0.39	Episodic Recall: 0.32	Deception: 0.34
Deception: 0.27	Cued Explicit Recogn: 0.26	Episodic Recall: 0.22
Fixation: 0.25	Imagined Obj/Scenes: 0.23	Film Viewing: 0.22

Notes: The number of estimated components from 2 to 20 components was explored. From 6 to 16 components, additional components emerged as subdivisions of lower order components, corresponding to a nested ontology. This table illustrates for 10 to 14 components, how the top tasks recruiting a cognitive component are re-distributed among the subdivided components when more cognitive components are estimated. For example, the 12-component estimate of C4 (^{12}C4) divided into the 13-component estimates of C4 (^{13}C4) and C5 (^{13}C5), while the remaining 11 components remained almost identical. As is evident from the table, the top 5 tasks for the ^{12}C4 component became the top two and three tasks that recruit ^{13}C4 and ^{13}C5, respectively. "(C)" and "(O)" indicate "covert" and "overt" respectively.

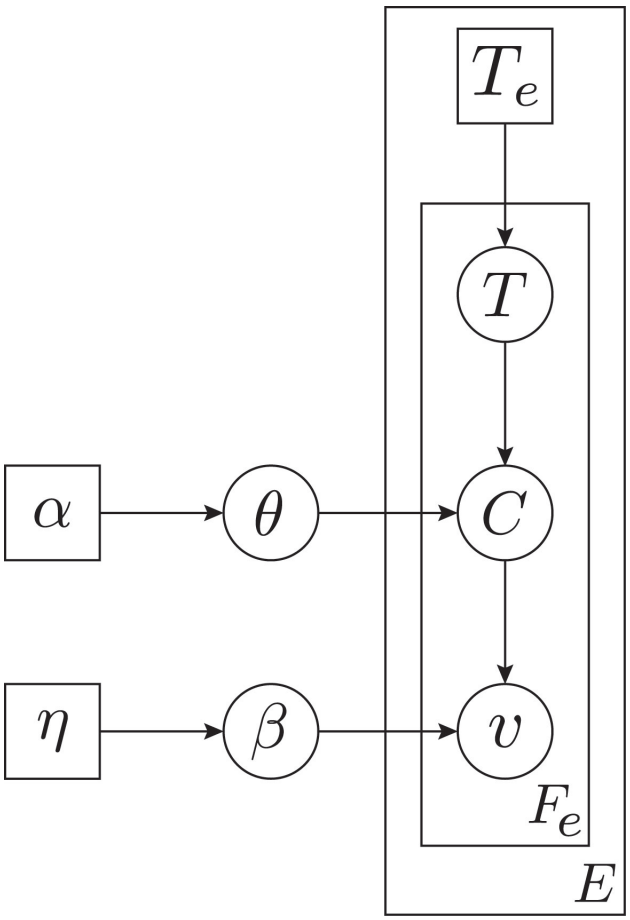

Figure S1. Formal mathematical representation of the author-topic model in the context of this work. This type of diagram is referred to as a graphical model (Blei et al., 2003). The circled variables represent random variables, while the squared nodes represent non-random parameters. The edges represent statistical dependencies. The model assumes a total of E experiments. The e-th experiment has F_e number of activated voxels and a set T_e of behavioral tasks. The f-th activated voxel has an observed location v, as well as a latent (unobserved) component C and latent (unobserved) task T associated with it. The variables at the corner of the rectangles indicate the number of times the variables inside the rectangles were replicated. Therefore T_e was replicated E times, once for each experiment. For the e-th experiment, the variables v, C and T were replicated F_e times, once for each activated voxel in the binary activation image. θ denote Pr(component | task) and β denote Pr(voxel | component). Therefore θ and β are matrices, where each row is a categorical distribution summing to one. α and η are hyperparameters parameterizing the Dirichlet priors on θ and β respectively.

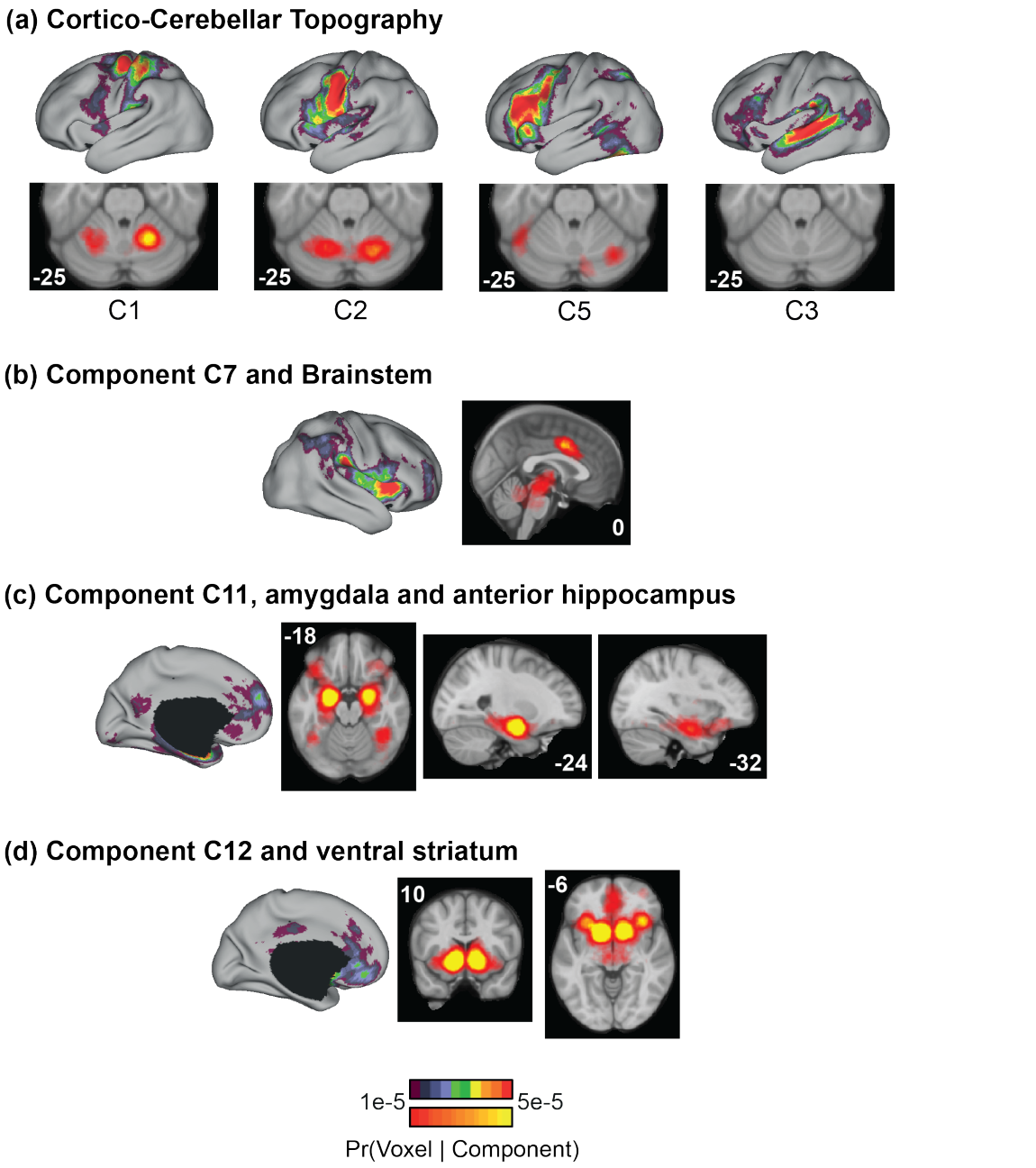

Figure S2. Probability of components activating subcortical structures Pr(voxel | component). (a) Cortico-cerebellar topography. Components C1, C2 and C5 activated distinct cerebellar territories from anterior to posterior, consistent with known cerebellar organization (Stooley and Schmahmann, 2009). Component C3 primarily activated auditory cortex, but not the cerebellum. (b) Component C7 activated the brainstem with high probability. (c) Component C11 activated the anterior hippocampus and amygdala with high probability. (d) Component C12 activated the ventral striatum with high probability. The top color bar represents the surface-based visualization of Pr(voxel | component), while the bottom color bar represents the volumetric slices highlighting subcortical structures.

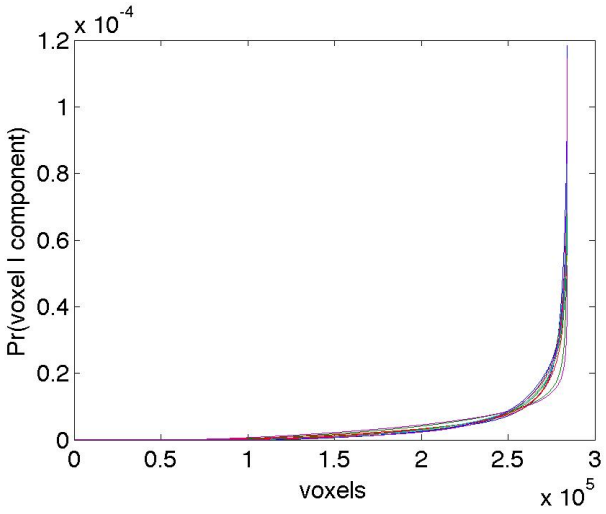

Figure S3. No Pr(voxel | component) is particularly diffuse. Figure plots the values of Pr(voxel | component) for each component, sorted from low to high and plotted. Each curve corresponds to one component. The curves are all tightly clustered suggesting that no component is especially spatially extensive.

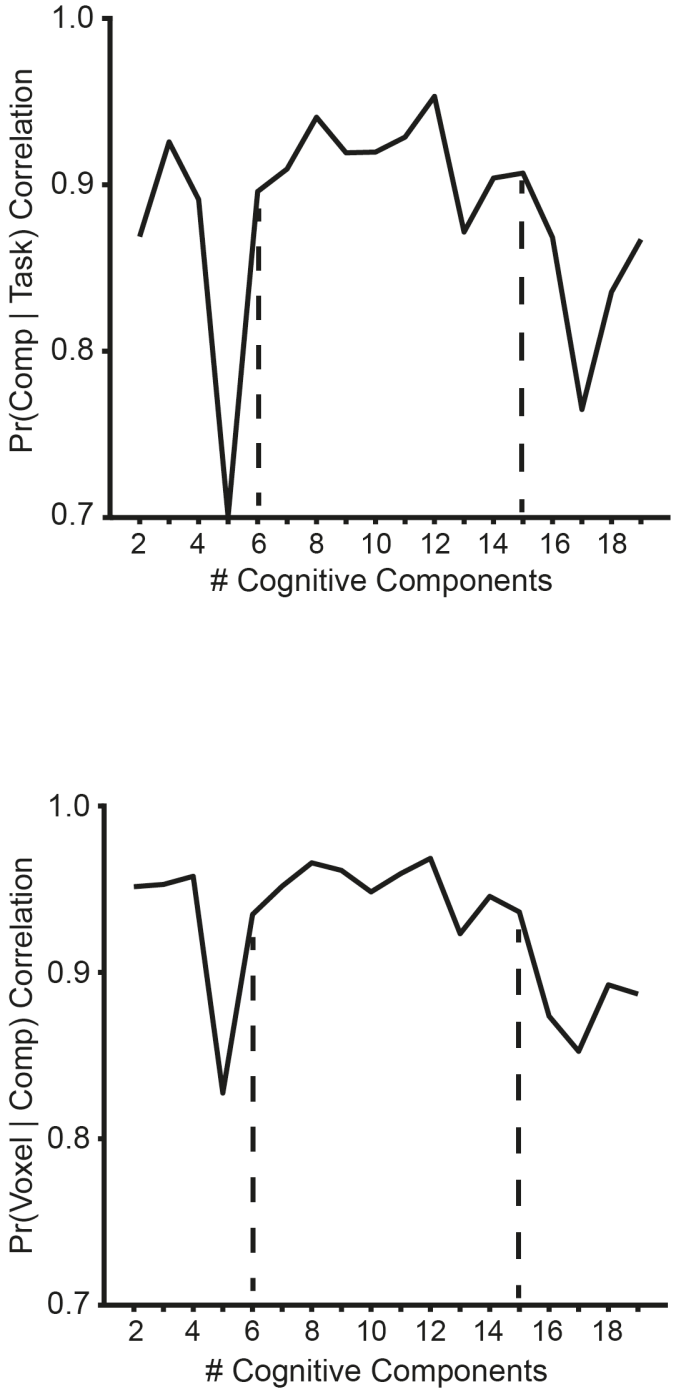

Figure S4. Quantification of nested ontology. A high correlation value at 'n' on the x-axis indicates a good quality of split from the n-component estimate to the (n+1)-component estimate. For instance, the high correlation value at 6 (dashed line) implies that the hypothesized split from 6-component to 7-component is good. Similarly, the high correlation value at 15 (dashed line) implies that the hypothesized split from 15-component to 16-component is good. Overall, the high correlation values from 6 to 15 indicate evidence of a nested ontology from 6 to 16 components. Quantifications are shown for Pr(Component | Task) *(top)* and Pr(Voxel | Component) *(bottom)*.

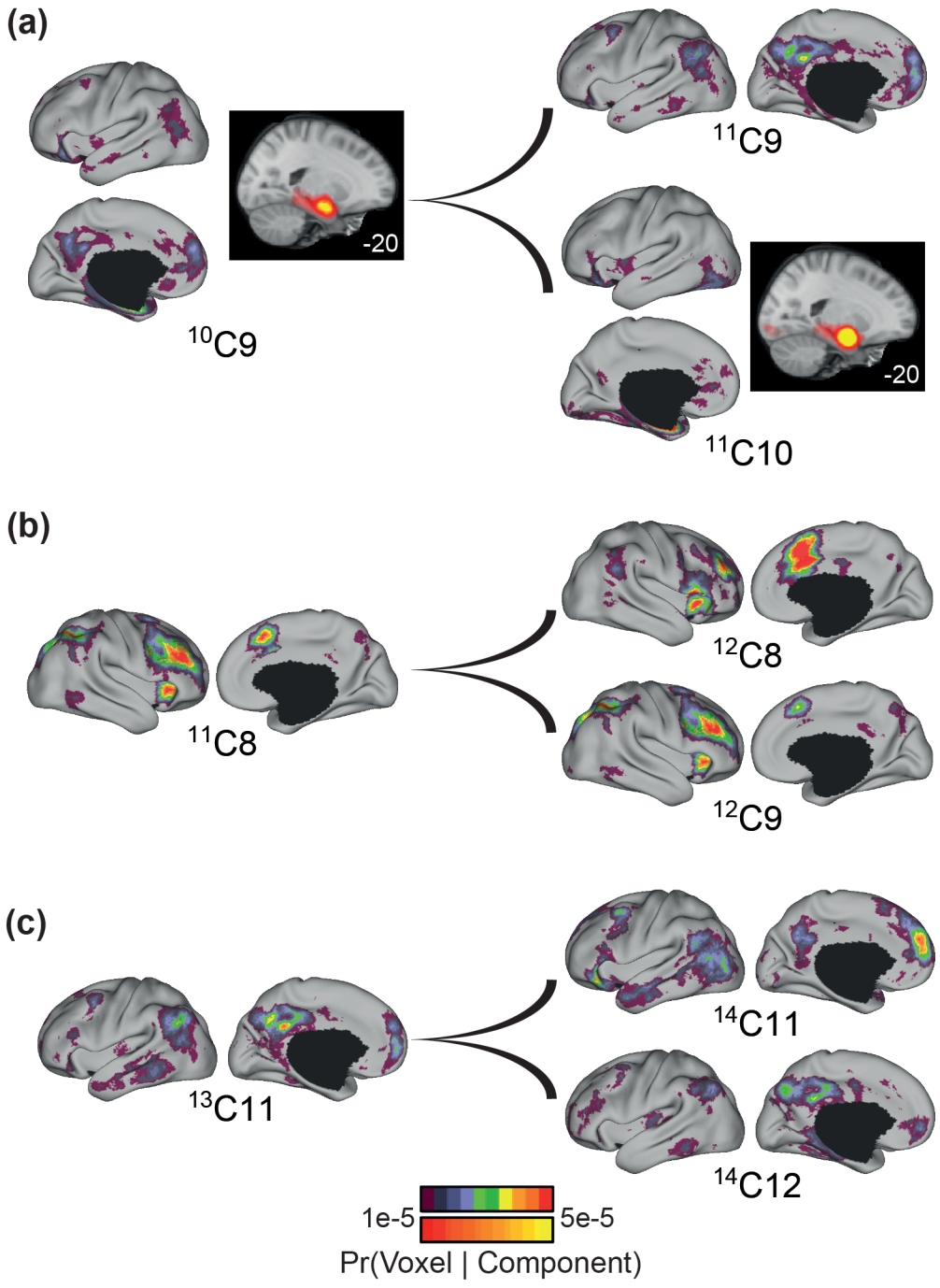

Figure S5. Nested ontology of Pr(voxel | component). Additional components emerged as subdivisions of lower order components, corresponding to a nested ontology. This figure illustrates how the voxels likely to be activated by a cognitive component are redistributed among the subdivided components as the number of components is increased from (a) 10 to 11, (b) 11 to 12 and (c) 13 to 14.

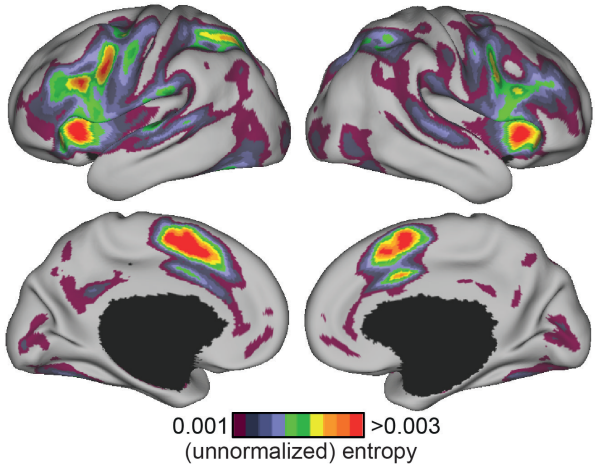

Figure S6. Alternative flexibility measure: unnormalized entropy of Pr(voxel | component). This figure shows an alternative functional flexibility measure, unnormalized entropy, defined as $-\sum_i Pr(voxel \mid component\ i) * log\ Pr(voxel \mid component\ i)$. The resulting functional flexibility map is similar to the original definition (Figure 5), with Pearson's correlation equal to 0.88.

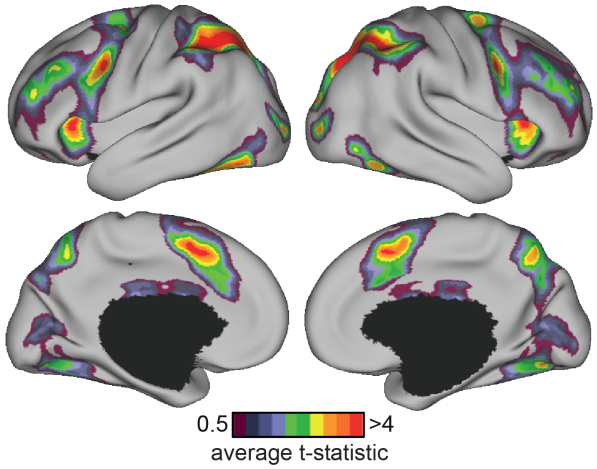

Figure S7. Multiple-Demand (MD) system estimated in Fedorenko et al. (2013). The overlay corresponds to t-statistic from multiple task contrasts that isolate cognitive demand processes as detailed in Fedorenko et al (2013), averaged across seven tasks and shown for both hemispheres. The unthresholded data in volumetric space can be found at http://imaging.mrc-cbu.cam.ac.uk/imaging/MDsystem.

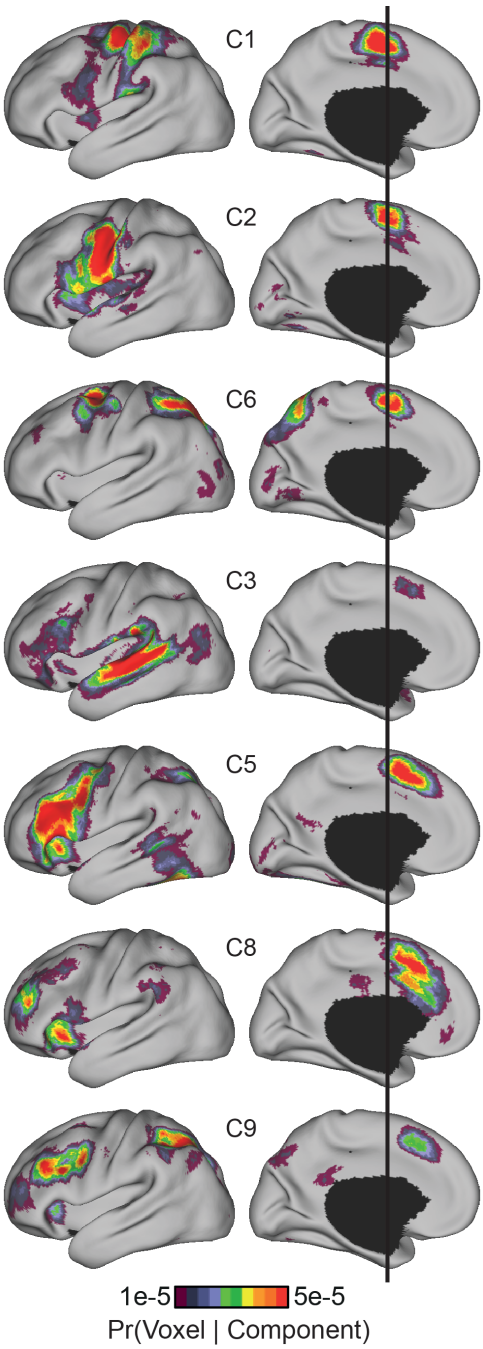

Figure S8. Possible functional gradient within posterior medial frontal cortex. Seven components from the 12-component estimate have high probability of activating the posterior medial frontal cortex. The components are arranged so that their activations within this brain region were roughly ordered posterior to anterior. The black vertical line provides a reference for comparing activation locations across components.

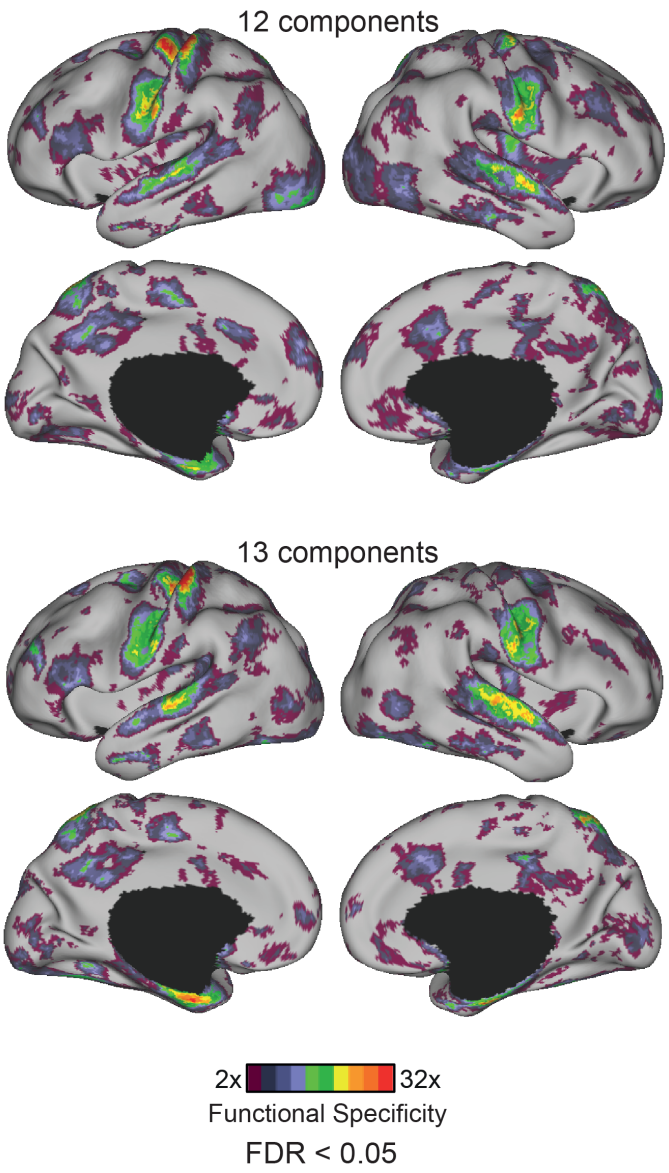

Figure S9. Functional specificity in the cerebral cortex for 12-component and 13-component estimates. Only regions with statistically significant (corrected for multiple comparisons for the entire cerebral cortex, FDR $q < 0.05$) functional specificity of at least two are shown. The somato-motor and auditory cortices exhibited higher functional specificity than the association cortex. Nevertheless, multiple components exhibited significant specificity in the association cortex, suggesting a fair degree of functional segregation in association cortex. Functional specificity estimates are similar across the 12-component and 13-component estimates: the Pearson's correlation between the two maps is 0.76. Differences in functional specificity estimates between 12-component and 13-component estimates arise mostly in the visual cortex. This is because of the division of the visual component into dorsal and ventral visual streams as the number of components increases from 12 to 13. This illustrates the scale-dependent nature of functional specialization. Note that the colorscale is logarithmic.

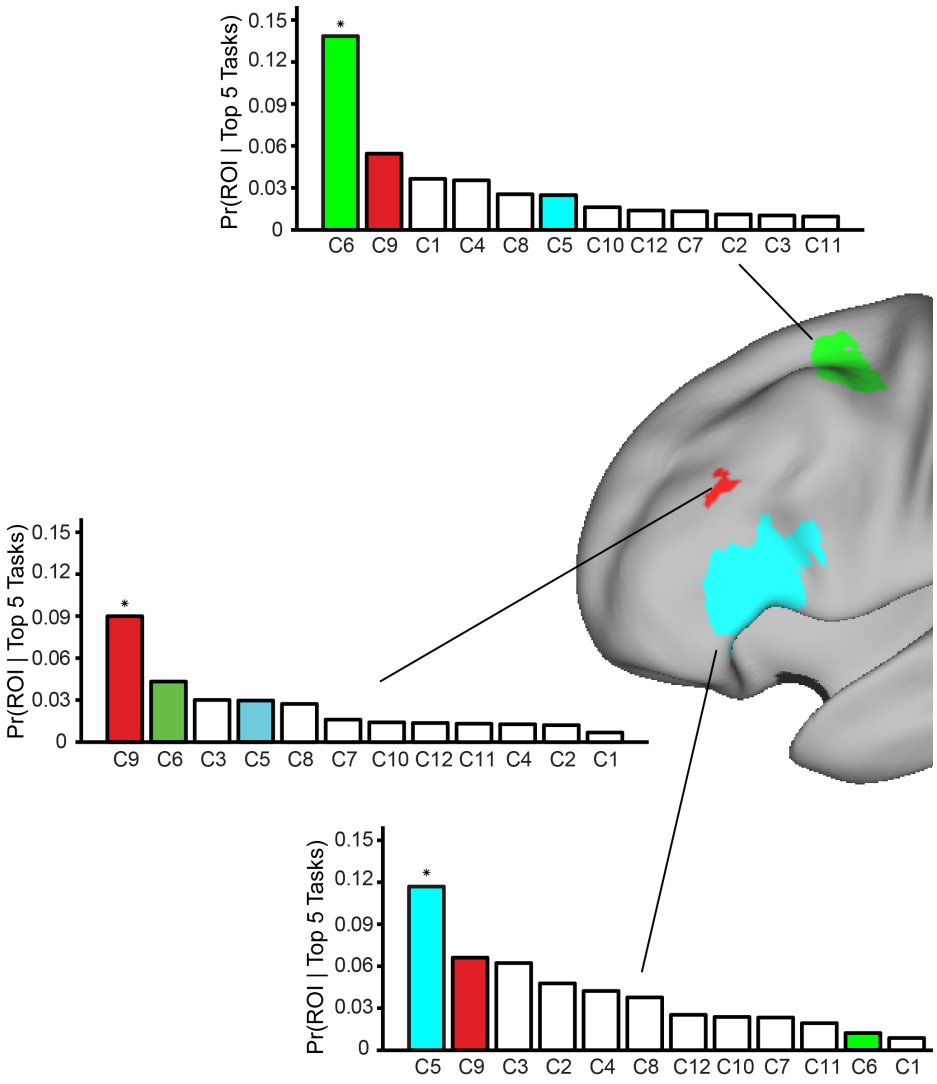

Figure S10. Functional specificity estimates are robust to analysis choices. This figure illustrates an alternative analysis for 3 of the 41 functionally specialized islands in lateral frontal and parietal cortices. For each island, the probability that the top five tasks of each component would activate the island was computed. The asterisks and colors indicate the most likely components as identified in the quantitative functional specificity maps (Figure 7). In the examples shown here, the top five tasks of the most likely component (as identified in Figure 7 and Table 1) had the highest probability of activating the respective islands. The agreement between this alternative analysis and the original functional specificity estimates ($p < 1e\text{-}5$) suggests the estimates truly reflected the BrainMap data, rather than being artifacts of the particular model or estimation procedure.