Web Appendix B: Additional information and supplementary figures and tables for Section 6: Applications

Additional information on the gender gene-set analysis

This data was originally analyzed via the GSEA method from Subramanian et al. (2005), and was later analyzed with a different method, which used t-tests, in Irizarry et al. (2009). We excluded 40 of the original 212 sets, in order to obtain nonoverlapping atoms. The gene-level statistics used in the EB estimation represented the signal-to-noise ratio (calculated in the original manuscript) and were standardized to have mean 0 and variance 1.

Additional information on the brain ROI analysis

A common problem in fMRI is the analysis of so-called group contrast data (Friston et al., 2007). Here, parameters from a first-stage subject-specific regression analysis are compared across subjects in standardized space. Each data point is a map of regression coefficients, conceptually representing a subject's blood oxygen level dependent (BOLD) response to an experimental stimulus. The test considered group level activation, that is, the areas of common response to the task across subjects, implying a one sample voxel-by-voxel t-test applied to the contrast data. The data from Henson et al. (2002), along with further background information and details on preprocessing, are freely available from the Statistical Parametric Mapping (SPM) website (http://www.fil.ion.ucl.ac.uk/spm/). The data are reanalyzed here, kindly having been given permission by the creators of SPM. As the anatomical atlas considered was at a finer resolution than the observed data, we down-sampled it to the lower resolution. This leads to small issues at boundary voxels of the parcellation that were not consequential for the overall analysis by impacting a very small percentage of tests.

Supplementary figures and tables

[Table 1 about here.]

[Figure 1 about here.]

[Table 2 about here.]

[Figure 2 about here.]

References

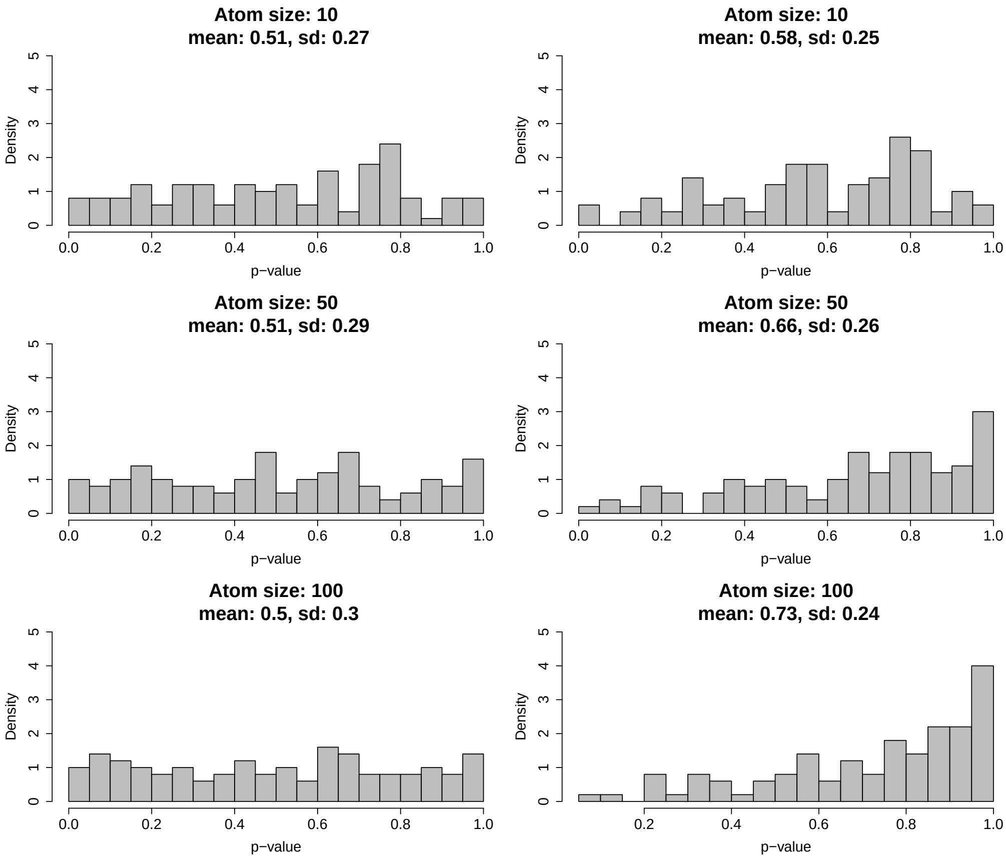
Friston, K., Ashburner, J., Stefan, K., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping The Analysis of Functional Brain Images.* Academic Press.

Henson, R. N. A., Shallice, T., Gorno-Tempini, M. L., and Dolan, R. J. (2002). Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex* **12,** 178.

Irizarry, R. A., Wang, C., Zhou, Y., and Speed, T. P. (2009). Gene set enrichment analysis made simple. *Statistical Methods in Medical Research* **18,** 565–575.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550.

Figure B1: Histograms of the p-values obtained from the `roast` method (Subfigure a) and `romer` method (Subfigure b). for atoms of sizes 10, 50, and 100 which have only null variables. Note as the set size increases, the histograms for the `romer` method are increasingly skewed towards 1.



(a) (b)

Figure B2: Histograms of $1-\widehat{afdr}$ (left panel) and q-values obtained from the `limma` method (right panel) for a fMRI brain dataset (Henson et al., 2002). Note that $1-\widehat{afdr}$ covers a much wider number of values than the q-values. Of the 117 regions of interest (ROIs), 51 have q-values less than 0.1 and 53 have q-values larger than 0.9.

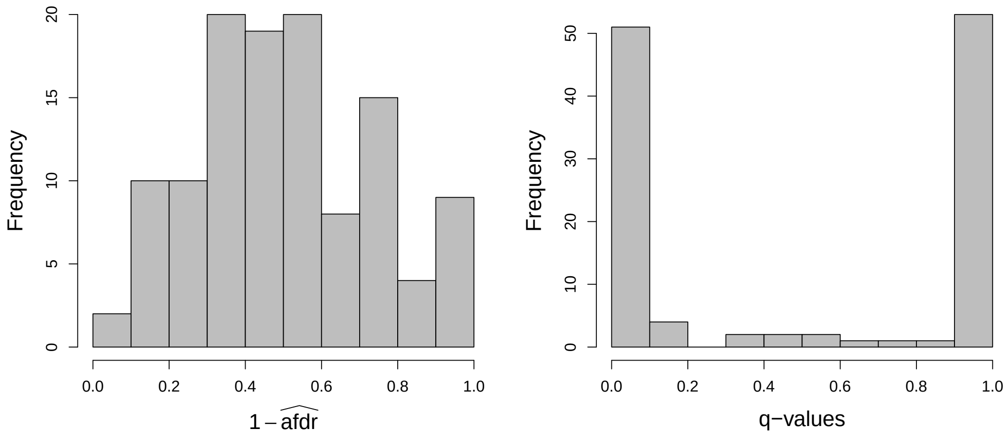


Table B1: Comparison of $1 - \widehat{afdr}$ for different fractions of alternatives and different sample sizes. The sample size if $4 \times N$, with $2N$ cases and $2N$ controls, resulting in variances of $1/N$ for each variable (all variables are normally distributed, with the null variables having mean 0 and the non-null variables having mean 1). Means and standard deviations are given for 100 simulation runs.

Fraction of alternatives	N=5	N=10	N=15	N=25	N=50
0	0.022 (0.012)	0.015 (0.010)	0.009 (0.009)	0.006 (0.006)	0.003 (0.004)
0.1	0.062 (0.019)	0.081 (0.019)	0.093 (0.015)	0.102 (0.010)	0.103 (0.005)
0.3	0.147 (0.030)	0.228 (0.027)	0.266 (0.020)	0.294 (0.012)	0.301 (0.004)
0.5	0.229 (0.037)	0.368 (0.030)	0.438 (0.025)	0.486 (0.015)	0.500 (0.004)
0.7	0.302 (0.041)	0.502 (0.040)	0.604 (0.032)	0.676 (0.017)	0.699 (0.004)
0.9	0.387 (0.054)	0.645 (0.048)	0.772 (0.034)	0.870 (0.015)	0.897 (0.005)

Table B2: Comparison of our method to the GSEA, t-test, and Wilcoxon rank test methods on a dataset from Subramanian et al. (2005). The top 5 sets according to $1 - \widehat{afdr}$ are shown.

Set	$1 - \widehat{afdr}$	`roast`			`romer`	
		p-value	q-value	estimated active proportion	p-value	q-value
chrYq11	0.398	0.001	0.172	0.562	0.002	0.258
chr3q13	0.041	0.626	1	0.182	0.609	1
chr12p12	0.039	0.651	1	0.100	0.445	1
chr7p22	0.038	0.896	1	0.087	0.776	1
chr9q21	0.036	0.650	1	0.086	0.330	1

Set	$1 - \widehat{afdr}$	GSEA		t-test	
		p-value	q-value	p-value	q-value
chrYq11	0.398	<0.001	<0.001	<0.001	<0.001
chr3q13	0.041	0.717	1	0.972	0.995
chr12p12	0.039	0.636	1	0.893	1
chr7p22	0.038	0.891	1	0.986	0.992
chr9q21	0.036	0.035	1	0.058	0.67