HCP beta-release of the Functional Connectivity MegaTrawl

April 2015 “HCP500-MegaTrawl” release

Stephen Smith[1] Diego Vidaurre[2] Matthew Glasser[3] Anderson Winkler[1] Paul McCarthy[1] Emma Robinson[1] Xu Chen[4] William Horton[3] Mark Jenkinson[1] Eugene Duff[1] Christian Beckmann[5] Mark Woolrich[2] Daniel Marcus[3] Deanna Barch[3] Kamil Ugurbil[6] Thomas Nichols[4] David Van Essen[3]

[1]FMRIB, Oxford University, Oxford, UK [2]OHBA, Oxford University, Oxford, UK [3]Washington University, St Louis, MO, USA [4]University of Warwick, Coventry, UK
[5]Donders Institute, Radboud University, Nijmegen, Netherlands [6]CMRR, University of Minnesota, Minneapolis, MN, USA

Summary

We summarise here an analysis of the relationships between imaging and non-imaging measures in the Human Connectome Project (HCP) [Van Essen 2013], using multivariate-prediction and univariate-regression to relate 187 non-imaging behavioural and demographic “subject measures” (SMs: age, sex, education, tobacco intake, IQ, reading ability, etc.) to 461 individuals’ functional connectivity data.

Resting-state fMRI (rfMRI) data from the 500-subjects 2014 HCP data release (HCP500-PTN) was processed, using data from all 461 subjects having four complete rfMRI runs. Using this cohort, network matrices from time-series partial correlations were used to predict (and regress against) 187 “subject-measure” (SM) variables, such as IQ, age and sex.

In addition to the descriptive text and citations below, which may be useful when writing papers that take advantage of these “higher-level” HCP outputs, please remember to include the generic HCP acknowledgements (and core HCP citations) when using HCP data in your research: see http://www.humanconnectome.org/documentation/citations.html

Subject-wise network matrix generation

Group-ICA [Smith 2014a, FastICA/MELODIC] was applied at 5 dimensionalities (d=25,50,100,200,300) to preprocessed rfMRI data, with surface-based alignment (“MSM-sulc”) utilising folding patterns to align different subjects’ surfaces with each other [Glasser 2013, Salimi-Khorshidi 2014, Robinson 2014]. These group-ICA “parcellations” (where each ICA component comprises a “node” or “parcel”) were used to estimate associated timeseries from each subject, with two methods: multiple-spatial-regression against the group-ICA spatial maps (“ts2”) or via eigenregression (“ts3”, [Smith 2014b]). From node timeseries, network matrices (netmats) were estimated using partial correlation with limited L2 regularisation (“NetmatMethod=3”, setting rho=0.01 in the Ridge Regression netmats option in FSLNets). This Parcels, Timeseries and Netmats (PTN) dataset (and associated documentation) is available at db.humanconnectome.org/data/projects/HCP_500 – as is the MegaTrawl.

MegaTrawl Methods

The MegaTrawl uses these 10 sets of netmats (5 dimensionalities x 2 regression methods) to relate functional connectivity to non-imaging SMs. For each SM we used two approaches:

1) **SM prediction** uses a *subjects X edges* matrix in multiple-regression-based prediction (of a given SM). Each edge is correlated with the SM; the weakest 50% are discarded. Elastic-net regularisation is then used (see below), with hyperparameters and number of edges optimised by 10-fold cross-validation; a further 10-fold outer-loop is used for final prediction. Family structure is taken into account by ensuring that no families are split across left-in/out groups. Prediction takes place after removing brain & head size (as estimated by FreeSurfer), overall head motion (a summation over all timepoints of timepoint-to-timepoint relative head motion) and acquisition date as confounds (the last of these is actually the “acquisition quarter”, which is useful to include because there was a slight change in rfMRI reconstruction code during the third acquisition year-quarter; in future we will instead use the actual reconstruction code version as the confound). To avoid the risk of the deconfounding process spreading “information” from training into testing data subsets (which would invalidate the leave-N-out strategy), for each fold of the leave-N-out loops, the deconfounding regression weights are estimated from just the training subset, and then later applied to deconfound the left-out subset. The outcome measure is “coefficient of determination”, CoD = 1-PredictionErrorVariance/SMVariance.

In order to cope with high dimensionality and enhance interpretability, we use Elastic net (EN) regularisation [Zou 2005]. EN, a combination of L1 and L2 regularisation, provides a sparse solution and balances the contribution of correlated edges, allowing one to identify all edges that are meaningfully related to the SM. For example, if we had a set of SM-relevant edges that are redundant relative to each other, EN, instead of dropping all but one, would weight their contributions and select all of them as a cluster. We apply EN regression in two steps: first, we apply multivariate feature selection using EN; second, we focus on prediction using EN only on the edges that were previously selected. This strategy improves the statistical efficiency of the estimation [Meinhausen 2007]. To solve the EN optimisation problem, we use a coordinate descent algorithm [Friedman 10], which can efficiently compute the regression parameters in an arbitrarily fine grid of regularisation parameter values.

2) **SM regression** fits a linear model for each netmat edge using the same confounds. Permutation testing is used to estimate familywise error-corrected p-values (two-tailed), i.e., corrected for the chance of one or more false positives over all edges; family structure is accounted for, using only permutations that do not break the complex cross-subject covariance structure [Winkler 2014]. We also report the number of edges that are significant at $p<0.05$ for comparison with the expected number (i.e., 5% of the number of netmat edges tested), with the (two-tailed uncorrected) p-values again derived from permutation testing.

The MegaTrawl also includes tests for network heritability [Chen 2014a, Chen 2014b] and representations of the group-mean parcellations/netmats, including interactive parcellation/connectome viewers [Lancaster,FSLNets].

Results summarised

Using empirical-null thresholding (i.e., mixture modelling applied to the deconfounded-space CoD values), sets of netmats with d<200 result in fewer significant predictions than expected by chance; d>=200 gives up to 60% more than expected by chance. Evaluated another way, Bonferroni correction of these empirical-null P-values over the 1870 (187x10) tests gives 11 significant predictions. Of 1870 regression tests, 301 identified 1 or more edges significantly correlated with an SM – 3 times more than expected by chance.

Overall, d=200 gave the largest number of strong predictions, with the multiple-spatial-regression node-timeseries-estimation method performing slightly better than eigen-regression. Here, the two *behavioural* SMs that almost reach significance after considering multiple testing are years of education (SM#10) and age-adjusted oral reading ability (SM#207).

In a separate analysis (not included in the MegaTrawl), we combined all 10 sets of netmats, along with146 FreeSurfer anatomical measures (regional cortical thickness and curvature averages) as well as 66 TBSS/ENIGMA regional FA averages from dMRI, for one big prediction/regression analysis (using the same analyses as described above). Performance was slightly worse than the best rfMRI results.

Web pages explained

The **Parcellation** link loads a web-based viewer ('Papaya' [Lancaster]) to allow MNI-space viewing of the high-dimensional group-ICA parcellation utilised for a given MegaTrawl analysis. The higher the dimensionality, the smaller are the individual "parcels" (i.e., the node maps are spatially less extended, and in some cases, have fewer non-contiguous regions). A semi-transparent overlay with multiple colours shows each ICA "parcel" in a different colour; although individual ICA maps are estimated (and, in general, used) without applying any thresholding, they are considered as binary maps for display purposes here – where each grey-matter voxel is colour-coded according to the index number of the ICA component having the highest value at that voxel. A separate overlay is also shown (solid red-yellow colouring), that shows one single unthresholded ICA component at a time. The index number is displayed at the top of the page, and this can be incremented/decremented using the "." and "," keys.

The group-ICA was carried out on data in grayordinates space (surface vertices and subcortical voxels [Glasser 2013]). For volumetric MNI-space display purposes here (and elsewhere in the MegaTrawl visualisations), grayordinate-space maps are written into MNI-space utilising an average cortical surface mapped onto 3D (scripts courtesy Saad Jbabdi).

The **group-average netmats** are visualised in 3 different ways. First, the raw group-averaged netmats are shown in a **matrix hierarchy** view, with full-correlation netmats shown below the diagonal and partial-correlation netmats above the diagonal. The nodes (rows and columns) have been reordered, bringing clusters of similarly-behaving nodes together (the hierarchical clustering is shown as part of the figure). Network edges are the elements of the netmat matrices, with red-yellow-brown depicting positive connectivity, and blue depicting negative connectivity. Partial correlation netmats are more interpretable than full correlation, and are used throughout the MegaTrawl in the SM predictions/correlations [Smith 2014b].

The second visualisation of the group netmats is **nodes+edges 1**. This is a different representation of the same partial correlation group netmat, focussing on nodes (thumbnail images) and edges. Red edges are positive correlations; blue are negative, and edge thickness represents connectivity strength. The major clusters of nodes are shown in separate pages, and then each separate node has its own page showing the strongest connections involving that node. Nodes within each node-sub-page are clickable, taking one to that node's own sub-page. For any given node's sub-page, other nodes are included if they connect to the primary node with an absolute connection strength greater than 75% of the strongest edge in the whole netmat. Similarly, edges between any pairs of nodes included are shown if they exceed the same threshold. These pages are created using the Graphviz package.

The third visualisation of the group netmats is **nodes+edges 2**. This is an interactive "circular" netmat representation where individual nodes can be clicked on to see their connections, with thresholding changeable by the user, along with various other aspects of the display. For a given node, a sub-plot can also be shown, giving the connections between just those nodes which connect to the original chosen node. This page was created by Paul McCarthy, using the D3 javascript library.

The heritability of entire netmats, nodes (combining across all a node's edges) and individual edges is shown on the **heritability calculations** page. This uses aggregate heritability methods [Chen 2014a,b]. The aggregate heritability for entire netmats is reported as a single number; this calculation relates to the boxplot that shows the pairwise-netmat-similarities for different kinds of pairs of subjects. Columns from netmats are then fed into similar calculations, generating aggregate heritabilities for each node. The node with the highest heritability is shown with a single thumbnail image. Finally, individual edges' heritabilities are calculated, and the 24 most heritable edges depicted as node-pairs. The text reports the number of "significant" edges' heritabilities, after correcting for multiple comparisons across all edges (though not correcting for the 10 MegaTrawl analyses). The coloured bars linking the two nodes in each pair reflects the group-average partial correlation edge, and the "value" quoted in the subplot title reflects the significance (1-P_{FWE}).

Finally, each SM (subject measure) then has its own **SM correlation/prediction** page. For many SMs, the netmat-based correlations/predictions are weak, and hence variable across the 10 MegaTrawl analyses. For SMs with stronger netmat predictability/correlation, results are more consistent across analyses.

First the **multivariate prediction** is shown. Here all netmat edges are used from each subject to attempt to get the best possible prediction of the SM. A scatterplot of the predicted SM value vs. the original SM value is shown, and the text reports the correlation (r) between the same. The coefficient-of-determination (percentage of variance explained by the prediction) is provided both for the original SM and after first regressing the confounds out of the SM and the netmats.

Secondly, the **univariate regression** of each netmat edge against the SM is shown. The N_{edges} regressions are tested for statistical significance, correcting for multiple comparisons across all edges (though not correcting for the 10 MegaTrawl analyses). The 24 edges with the strongest correlation with the SM are shown.

Conclusions

This "HCP MegaTrawl" provides a resource for researchers interested in relating brain connectivity to non-imaging subject measures. It may serve as a useful benchmark that capitalizes on state-of-the-art methodology. It also provides a salutary lesson in the consequences of over-fishing and multiple testing; indeed, the term "MegaTrawl" is intended to emphasise that large-scale exploratory analyses have the potential to "always find something". Using a range of parcellation dimensionalities and methods for estimating timeseries and network matrices, this MegaTrawl comprises over 2000 sets of tests. Although fewer than half of the final expected number of HCP subjects were used here, the number of significant relationships found to date is not high, suggesting that much work remains to be done to optimally mine the HCP dataset for significant relationships between behaviour and functional connectivity. For example, sensitivity to such relationships may be enhanced by improved intersubject registration based on areal features (myelin maps and resting-state fMRI; Robinson et al., 2014), as will be provided in a forthcoming HCP data release.

Acknowledgements

We are very grateful to Jack Lancaster and Michael Martinez for the Papaya tool (and for help with getting it working well for the MegaTrawl).

References

[Van Essen 2013] DC Van Essen. The WU-Minn Human Connectome Project: An overview. NeuroImage 2013.

[Smith 2014a] SM Smith. Group-PCA for very large fMRI datasets. NeuroImage 2014.

[Glasser 2013] MF Glasser. The minimal preprocessing pipelines for the Human Connectome Project. NeuroImage 2013.

[Salimi-Khorshidi 2014] G Salimi-Khorshidi. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. NeuroImage 2014.

[Robinson 2014] EC Robinson. MSM: A new flexible framework for Multimodal Surface Matching. NeuroImage 2014.

[Smith 2014b] SM Smith. Methods for network modelling from high quality rfMRI data. OHBM 2014.

[Zou 2005] H Zou. Regularization and variable selection via the elastic net. J Royal Statistical Society: Series B 2005.

[Friedman 2010] J Friedman, Regularization Paths for Generalized Linear models via Coordinate Descent, J Statistical Software 33 (2010)

[Meinhausen 2007] N Meinhausen. Relaxed Lasso. Computational Statistics & Data Analysis 52 (2007).

[Winkler 2014] A Winkler. Multi-level Block Permutation for the Human Connectome Project. OHBM 2014.

[Chen 2014a] X Chen. A method for fast whole-brain aggregate heritability estimation. OHBM 2014.

[Chen 2014b] X Chen. APACE: Accelerated Permutation Inference for the ACE Model. OHBM 2014.

[Lancaster] J Lancaster. Papaya: http://ric.uthscsa.edu/mango/papaya.html