Multiscale statistical testing for connectome-wide association studies in fMRI

Pierre Bellec\textsuperscript{a,b,*}, Yassine Benhajali\textsuperscript{a,c}, Felix Carbonell\textsuperscript{d}, Christian Dansereau\textsuperscript{a,b}, Zarrar Shehzad\textsuperscript{f,g}, Geneviève Albouy\textsuperscript{a,h}, Maxime Pelland\textsuperscript{b}, Cameron Craddock\textsuperscript{f,g}, Olivier Collignon\textsuperscript{b}, Julien Doyon\textsuperscript{a,h}, Emmanuel Stip\textsuperscript{i,j}, Pierre Orban\textsuperscript{a,i}

\textsuperscript{a}Functional Neuroimaging Unit, Centre de Recherche de l’Institut Universitaire de Gériatrie de Montréal
\textsuperscript{b}Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada
\textsuperscript{c}Department of Anthropology, University of Montreal, Montreal, Quebec, Canada
\textsuperscript{d}Biospective Incorporated, Montreal, Quebec, Canada
\textsuperscript{e}Department of Psychology, Yale University, New Haven, CT, United States of America
\textsuperscript{f}Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, United States of America
\textsuperscript{g}Center for the Developing Brain, Child Mind Institute, New York, NY, United States of America
\textsuperscript{h}Department of Psychology, University of Montreal, Montreal, Quebec, Canada
\textsuperscript{i}Department of Psychiatry, University of Montreal, Montreal, Quebec, Canada
\textsuperscript{j}Centre Hospitalier de l’Université de Montréal, Montreal, Quebec, Canada

Abstract

Alteration in brain connectivity, as captured by fMRI, has been found to be associated with a variety of clinical disorders. Yet, the systematic identification of connections associated with a disease state is a “needle in a haystack” problem. There are numerous connections to test and potentially very few of them show a strong association with the variable of interest. The number of brain parcels (or scale) used to measure connections is a key parameter, as more parcels means more anatomically accurate measures, but also means many more connections to test and thus more risk to find spurious associations by chance. This work presents a multiscale approach designed to assess the impact of scale on connectome-wide association studies. A mass univariate general linear model is applied independently at each connection, for functional networks generated through clustering at different scales. The multiple comparison problem is addressed at each scale independently, using the false-discovery rate. An omnibus permutation test is implemented to reject the global null hypothesis of no association across all scales. Our results support the fact that, when the global null is rejected, the false-discovery rate is well controlled both within and across scales. On simulations, we showed that multiscale group differences can be detected

*Corresponding author
Email address: pierre.bellec@criugm.qc.ca (Pierre Bellec)
with excellent sensitivity in large sample size \((n=100)\), and good sensitivity in moderate sample size \((n=40)\). On three real resting-state datasets (schizophrenia, congenital blindness, motor practice), our results presented an excellent face validity. In addition, the multiscale method performed better than current techniques such as network-based statistics or multivariate distance matrix regression, both on simulations and real datasets. Taken together, our findings demonstrate that our approach is a powerful tool to mine for associations across the whole connectome, at multiple scales.

**Keywords:** fmri, general linear model, functional parcellation, multiple comparison, false discovery rate, multiscale analysis, connectome

**Highlights**

- A new statistical method to test the association between phenotypes and functional connectivity at multiple scales (number of brain parcels).
- The false-discovery rate is controlled both within- and between scales, with state-of-the-art sensitivity.
- A “group” method to control the false-discovery rate in connectomes, with improved sensitivity to sparse signals compared to the standard Benjamini-Hochberg procedure.

1. **Introduction**

**Main Objective.** Estimates of brain connectivity derived from functional magnetic resonance imaging (fMRI) have been found to be associated with a wide variety of clinical disorders \(\text{[Castellanos et al.] 2013}\). Rather than focusing on a limited set of \textit{a priori} regions of interest, a recent trend is to perform statistical tests of association across the whole connectome, i.e., at every possible brain connection \(\text{[Shehzad et al.] 2014}\). Akin to genome-wide association studies, such connectome-wide association studies (CWAS) raise a large multiple comparison problem: with \(10^3\) brain parcels, there are about \(5 \times 10^5\) connections between parcels to test. When a massive number of univariate tests are performed, it becomes difficult to simultaneously reach acceptable rates of false positives (specificity) and false negatives (sensitivity). The number of multiple comparisons can be reduced by using fewer, bigger parcels, but then associations that are highly specific to small anatomical locations may be lost. In addition to these statistical concerns, examining and interpreting all findings in a connectome quickly becomes challenging in practice when the number of parcels gets large. The number of brain parcels (or scale) thus has critical implications both on the statistical power of a CWAS and the interpretability of the results. The main objective of this work was to develop a statistical framework able to perform a CWAS at multiple scales, instead of relying on a single parcellation, such that the impact of the spatial resolution could be assessed and optimized.
Mass-univariate connectome-wide association studies. The mass-univariate approach to CWAS (Worsley and Friston, 1995) consists of independently estimating a general linear model (GLM) at each connection, i.e. for any given pair of brain parcels. In the GLM, a series of equations are solved to find a linear mixture of explanatory variables (called covariates) that best fit the connectivity values observed across the many subjects. A $p$-value is generated for each connection to quantify the probability that the estimated strength of association between this connection and a covariate of interest could have randomly arisen in the absence of a true association. For a single test, a $p$-value below 0.05 is typically considered as strong evidence of an association, yet in a connectome, even in the absence of any true association, about 5% of connections would get reported at this significance level. Given the very large number of tested connections, this could result in 1,000s of false positive findings, unless an adequate correction for multiple comparisons is applied. Random field theory is routinely used to correct for multiple comparisons in 3D statistical parametric maps (SPMs), where a statistical test is performed at each voxel. Early efforts by Worsley et al. (1998) aimed at extending random field theory from a 3D brain activity manifold to a 6D connectome manifold - a 3D space connected to a 3D space. Random field theory offers a control of the family-wise error (FWE) rate, i.e. the probability of one or more false positive arising across the whole connectome. Unfortunately, the strict control of false positives offered by FWE in a very large space like a connectome generally comes at a cost of low sensitivity. As an alternative, Shehzad et al. (2014) recently proposed to run a multivariate test for each possible region-to-brain connectivity map, called multivariate distance matrix regression (MDMR). The MDMR approach effectively performs one test per parcel or voxel, instead of one test per connection, and thus shrinks the connectome multiple comparison problem back to a standard SPM. MDMR was designed to screen for promising seed-based connectivity maps worthy to explore in a subsequent, independent analysis, yet it does not provide valid statistics at the level of a single connection.

Parcellations and testing. A straightforward way to systematically test every possible connections without succumbing to multiple comparisons is to reduce the number of brain parcels. This can be achieved using a set of brain regions delineated based on anatomical landmarks such as the AAL template (Tzourio-Mazoyer et al., 2002), see for example the study of Wang et al. (2007) on Alzheimer’s disease. Rather than using anatomical parcels, it is possible to build data-driven functional brain parcels that optimally capture the underlying spatial correlation structure of the fMRI time series at the voxel level (Shen et al., 2010; Craddock et al., 2012; Blumensath et al., 2013). It has been shown that task activation detection with the GLM at the parcel level was more sensitive than voxelwise analysis (Lu et al., 2003; Thirion et al., 2006), and the same was reported for the accuracy of multivariate prediction (Xu et al., 2010). However, even with a relatively small number of brain parcels $N = 100$, there are still about 5000 multiple comparisons. The work of Zalesky et al. (2010a) thus proposed to use uncorrected threshold on the individual $p$-values, but then to
identify to which extent the connections that survive the test are interconnected. This extent measure is compared against what could be observed under a null hypothesis of no association, implemented through permutation testing. This approach, called Network-Based Statistics (NBS), is the connectome equivalent to the “cluster-level statistics” used in SPMs. As the authors noted themselves, the NBS only offers a loose control of false-positive rate at the level of a single connection, but can be used to reject the possibility that a group of significant findings could be observed by chance in the FWE sense.

**Multiscale parcellations.** Few investigators have examined how to select the parcels, and in particular their number, in order to maximize the statistical power of a GLM analysis. Work by Abou Elseoud et al. (2011) systematically explored the impact of scale (number of components) on the ability of a dual-regression ICA analysis to discriminate a group of patients suffering from non-medicated seasonal affective disorder compared to normal healthy controls. The authors concluded that the number of significant findings was maximized at scale 45 (in this case, 45 independent components). The impact of the number of brain parcels was also investigated at much higher scales (from 50 to 3000+) by Shehzad et al. (2014), who concluded in their case that the results of the association between resting-state and intelligence quotient was consistent across scales. It should be noted that, in the above-mentioned studies (Abou Elseoud et al., 2011; Shehzad et al., 2014), the authors did not investigate the implications that testing many scales would have in terms of the control of false positives. Higher scales may indeed result in greater rates of false positives, as more components also resulted in more multiple comparisons (Meskaldji et al., 2011). Interestingly, the issue of multiscale testing was raised two decades ago in the context of SPM and Gaussian smoothing, and a theoretical framework based on random field theory had been developed to control for FWE when testing at multiple scales, i.e. different sizes of the spatial smoothing kernel (Poline and Mazoyer, 1994; Worsley et al., 1996). Multiscale random field theory has however not been generalized to connectomes, to the best of our knowledge, and was also limited to Gaussian blurring kernels, rather than data-driven brain parcels. We are thus not currently aware of a valid statistical framework to examine the results of a CWAS with data-driven brain parcellations at multiple scales.

**Specific objectives.** In this paper, our main objective was to develop a novel statistical framework, called multiscale statistical parametric connectome (MSPC), to perform a CWAS at multiple spatial scales. A series of data-driven functional brain parcellations was first generated using a clustering approach (Bellec et al., 2010). These parcels were used to derive connectomes for a group of subjects, and this process was repeated using different numbers of parcels were investigated. A GLM was then applied on the connectomes and the false-discovery rate (FDR) was controlled at each scale. Two methods were implemented to control the FDR: the popular approach by Benjamini and Hochberg (1995) as well as a recent algorithm optimized for sparse signals (Hu et al., 2010). We developed an omnibus test to assess if the number of significant findings (called
discoveries) across scales could have been observed under the global null hypothesis of no association at any scale. Our key hypothesis was that, although the FDR is controlled independently at each scale, it is also well controlled across scales when the global null is rejected. Our primary objective was to support this hypothesis empirically. Our second objective was to evaluate the sensitivity of the MSPC method, and in particular to determine which variant of the FDR control algorithm was most sensitive. Our third and last objective was to compare the sensitivity and specificity of the MSPC method with recent detection techniques applied at a single scale, namely NBS (Zalesky et al., 2010a) and MDMR (Shenhad et al., 2014). We conducted a series of experiments involving both simulated and real datasets to address these specific objectives, which have been summarized, along with the main findings, in Table 1.

Simulations. We first generated simulation datasets by mixing real fMRI data with synthetic signal. Importantly, the simulation datasets were designed to have a well defined ground truth in terms of true and false positives when using a multiscale set of functional parcels. We simulated group differences using a variety of scenarios covering: (1) different effect sizes; (2) different amounts of true positives; (3) different number of subjects per group. As part of the simulation study, we also performed CWAS on real fMRI datasets using random covariates of interest, thus providing insights in the behaviour of the methods under the global null hypothesis, where there is no true association to find.

Real datasets. We evaluated the MSPC, NBS and MDMR on three real datasets: (1) a study comparing patients suffering from schizophrenia with healthy control subjects (referred to here as the SCHIZO dataset, with a large sample size, $n = 146$); (2) a study on patients suffering of congenital blindness, compared to sighted controls (referred to here as the BLIND dataset, with a small sample size, $n = 31$), and; (3) a study of motor learning, where resting-state data were acquired before and after learning of a new motor task (referred to here as the MOTOR dataset, with a moderate sample size, $n = 54$). These three datasets were chosen to assess if MSPC would be able to uncover statistically significant effects in a variety of sample sizes. Also, we had strong a priori hypotheses for those three analyses: changes in the visual network for the BLIND dataset, in the motor network for the MOTOR dataset and a more general dysconnectivity in the SCHIZO dataset. These a priori were useful to qualitatively assess the face validity of the technique.
<table>
<thead>
<tr>
<th>Objective</th>
<th>Experiment(s)</th>
<th>Finding(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop a statistical framework for multiscale GLM analysis of connectomes.</td>
<td>All relevant algorithms are described in Section 2.</td>
<td>Not applicable.</td>
</tr>
<tr>
<td>Assess the specificity of MSPC in the absence of signal.</td>
<td>“Negative control experiments” where random subgroups are compared in the Cambridge sample.</td>
<td>The FWE under the global null is controlled at nominal level by the permutation test (Figure 5).</td>
</tr>
<tr>
<td>Assess the specificity of MSPC within and across scales.</td>
<td>Multiscale simulation of group differences in the Cambridge sample.</td>
<td>When including a permutation test for the global null, the FDR is controlled at nominal level both within and across scales (Figure 3 and Figure 6). The FDR at the peak of discoveries is slightly liberal (Figure 6).</td>
</tr>
<tr>
<td>Assess the sensitivity of MSPC, across scales or at the peak of discovery.</td>
<td>Multiscale simulation of group differences in the Cambridge sample.</td>
<td>The sensitivity across scales was excellent for a large sample size, n = 100, and good for moderate sample size, n = 40, as long as there were a large effect size and many true positives (Figure 4 and Figure 6). The highest sensitivity was obtained by looking at the scale with peak percentage of discovery.</td>
</tr>
<tr>
<td>Assess the face validity of the results identified with MSPC on real data.</td>
<td>Three datasets were analyzed: BLINDS, SCHIZO and MOTOR. Strong a priori hypotheses on the networks involved in each contrast were available.</td>
<td>The MSPC identified plausible changes in connectivity in all three analyses (Figure 8).</td>
</tr>
<tr>
<td>Compare two algorithms controlling for the FDR: classical BH and group FDR.</td>
<td>Multiscale simulation in the Cambridge dataset, and comparison on the three real datasets.</td>
<td>The group FDR is more sensitive across scales (Figure 3 and Figure 8).</td>
</tr>
<tr>
<td>Compare the MSPC with NBS and MDMR.</td>
<td>Multiscale simulation in the Cambridge dataset, and comparison on the three real datasets.</td>
<td>The MSPC with group FDR had overall the best sensitivity (Figures 6 and 8).</td>
</tr>
<tr>
<td>Assess how multiscale analysis can contribute to interpret results.</td>
<td>Exploration of connectivity changes in SCHIZO.</td>
<td>High scales clarified the spatial distribution of changes in connectivity seen at low scales.</td>
</tr>
</tbody>
</table>

Table 1: Summary of the objectives, experiments and findings of the paper.
2. Statistical testing procedures

2.1. Functional parcellations

The first step to build a connectome is to select a parcellation of the brain, with \( R \) networks. In this work, we used functional brain parcellations, aimed at defining groups of brain regions with homogeneous time series. A number of algorithms have been proposed with additional spatial constraints, to ensure that the resulting parcels are spatially connected [Lu et al., 2003; Thirion et al., 2006; Craddock et al., 2012]. However, from a pure functional viewpoint, the spatial constraint seems somewhat arbitrary, as functional units in the brain at low resolution encompass distributed networks of brain regions with homotopic regions often being part of a single parcel [De Luca et al., 2006; Damoiseaux et al., 2006]. Some works have thus used distributed parcels as the spatial units to measure functional brain connectivity, e.g. [Jafri et al., 2007; Marrelec et al., 2008]. To generate the functional parcellations, we relied on a recent method called “Bootstrap Analysis of Stable Clusters” (BASC), which can identify consistent functional parcels for a group of subjects [Bellec et al., 2010], using a hierarchical cluster with Ward’s criterion both at the individual and the group levels. The functional parcels can be generated at any arbitrary resolution (within the range of the fMRI resolution), and we considered only clusters generated at the group level. So, in this work, the word scale refers to the number of clusters as generated by the group BASC procedure, and each scale includes includes non-overlapping networks of brain regions that are not necessarily spatially contiguous.

2.2. Functional connectome

For each scale, and each pair of distinct networks \( i \) and \( j \) at this scale, the between-network connectivity \( y_{i,j} \) is measured by the Fisher transform of the Pearson’s correlation between the average time series of the networks. The within-network connectivity \( y_{i,i} \) is the Fisher transform of the average correlation between time series of every pair of distinct voxels inside network \( i \). The connectome \( Y = (y_{i,j})_{i,j=1}^{R} \) is thus a \( R \times R \) matrix. Each column \( j \) (or row, as the matrix is symmetric) codes for the connectivity between network \( j \) and all other brain networks, or in other word is a full brain functional connectivity map. See Figure 1a-b for a representation of a parcellation and associated connectome. Connectomes are generated independently at each scale. See Appendix A for a more formal description of the connectome generation.

2.3. Statistical parametric connectome

For a scale with \( R \) parcels, there are exactly \( L = R(R+1)/2 \) distinct elements in an individual connectome \( Y \). This connectome can be stored as a \( 1 \times L \) vector, where the brain connections have been ordered arbitrarily along one dimension. When functional data is available on \( N \) subjects, the group of connectomes is then assembled into a \( N \times L \) array \( Y = (y_{n,l}) \), where \( n = 1, \ldots, N \) each code for one subject and \( l = 1, \ldots, L \) each code for one connection. A general linear
Figure 1: **General linear model applied to connectomes.** The connectivity is measured between $R$ brain parcels generated through a clustering algorithm (panel a). The connectome is a $R \times R$ matrix measuring functional connectivity between- and within-networks. See main text for the definition of the connectivity measures. Note that each column of the connectome is equivalent to a full brain functional connectivity map for a particular seed region (panel b). The association between phenotypes and connectomes is tested independently at each connection using a general linear model at the group level (panel c). The results presented here are for illustration purpose only, and not related to the results presented in the application sections of the manuscript.

model (GLM) framework can then be used to test the association between brain connectivity and a trait of interest, such as the age or sex of participants. All of these $C$ explanatory variables are entered in a $N \times C$ matrix $X$. The variables are typically corrected to have a zero mean across subjects, and an intercept (i.e. a column filled with 1) is added to $X$. The GLM relies on the following stochastic model:

$$Y = XB + E,$$

with the following matrices:

- $Y$ is a $N \times L$ matrix where each row codes for a subject, and each column codes for a connection,
- $X$ is a $N \times C$ matrix of explanatory variables (or covariates) where each row codes for a subject and each column codes for a covariate,
• \( B \) is an unknown \( C \times L \) matrix of linear regression coefficients where each row codes for a covariate and each column codes for a connection,

• \( E \) is a \( N \times L \) random (noise) multivariate Gaussian variable, with similar coding to \( Y \).

As the data generated from different subjects are statistically independent, and under an homoscedasticity assumption, the regression coefficients \( B \) can be estimated with ordinary least squares. For a given “contrast” vector \( C \) of size \( 1 \times C \), the significance of \( C.B \) can be tested with a connectome of \( t \)-test \((t_l)^{L}_{l=1}\), with associated \( p \)-values \((p_l)^{L}_{l=1}\). The quantity \( p_l \) controls for the risk of false positive findings at each connection \( l \). The GLM applied on connectomes is illustrated in Figure 1c. See Appendix B for the equations related to the estimation and testing of regression coefficients in the GLM.

2.4. The Benjamini-Hochberg FDR procedure

In traditional (univariate) applications, a \( p \) value below 0.05 is considered as a solid evidence of an effect. For a given scale, there are however a large number \( L \) of such tests on \( p_l \) to perform. So even with a 5% chance of a false positive at connection \( l \), in the absence of any effect one would report 5% of the whole connectome as associated with a significant effect. The significance value applied on \( p_l \) thus needs to be adjusted for this multiple comparison problem. We implemented two strategies for controlling the FDR. The idea behind the FDR is not to strictly control the probability to observe at least one false positive (a quantity know as family-wise error, FWE), but rather to control the proportion of false positive amongst the findings. Note that controlling for the FDR is not necessarily a more liberal attitude than controlling for the FWE; if the global null hypothesis is verified, i.e. all discoveries are false positive, then the FDR is exactly the FWE. In the presence of true discoveries, however, the FDR procedure tolerates in general more noise than a FWE approach. The actual number of false discoveries will dependend on the amount of signal (true positive) present in the data, and is therefore a context-dependent question. The popular Benjamini-Hochberg (BH) procedure was proposed to control the FDR \( q \) to an acceptable level \( \alpha \) (Benjamini and Hochberg, 1995). This algorithm was designed for independent tests, but it has been shown to have a satisfactory behaviour even in the presence of positive correlation between the tests \( p_l \) (Benjamini and Yekutieli, 2001).

2.5. The group FDR procedure

A documented limitation of the BH procedure is its lack of sensitivity to sparse signals (Hu et al., 2010). If only a small subset of brain parcels show some significant effects, the maps of low signal will contribute a lot of noise and the FDR threshold that is set globally will be quite conservative. A connectome has a natural structure: each column represents a full brain connectivity map, where the parcel associated with the column acts like the seed (see Figure 1h). Some regions of the brain are likely to be particularly impacted by the
trait of interest, and be associated with many discoveries, while others will be associated with very few or even no discoveries. This means that some columns in the connectome will be associated with many more discoveries than others. These columns thus represent natural groups of tests within the full connectome, yet this structure is completely ignored in the BH procedure. Several solutions have recently been proposed for this issue (Efron 2008; Cai and Sun 2009; Hu et al. 2010; Benjamini and Bogomolov 2011). We focused in this work on the least-slope variant of the group FDR procedure proposed by Hu et al. (2010). The idea of the group FDR is to first assess the presence of signal in each column (map) independently, and then re-weight the original p-values based on the presence (or absence) of promising signal. It was shown that the group FDR procedure controls asymptotically the global FDR for independent tests, yet they are expected to be much more sensitive than BH for sparse signals (Hu et al.) (2010). The group FDR was also reported to behave correctly in the presence of positive dependency between tests, although no analytical properties were derived in that case. We investigated experimentally the differences between the BH and group FDR procedures, both on simulated data (Section 3) and real data (Section 4). The Appendix C provides a formal description of both the global and the group FDR algorithms.

2.6. Multiscale testing

To explore the association of connectomes and a trait of interest at multiple scales, a multiscale clustering decomposition was first generated with BASC, and then the GLM was estimated with FDR control at each scale independently, see Figure 2. This introduces a new level of multiple comparisons, this time across scales rather than across connections. There is however a qualitative difference between the multiple comparison problems encountered across multiple connections, and across multiple scales. In a mass univariate GLM connectome analysis, the possibility exists of a sparse signal, i.e. some connections show an effect, while others do not. In a multiscale analysis, a detection at any scale essentially implies the presence of signal at all the other scales. In the presence of signal, i.e. with true positive present at every scale, Efron (2005) suggested that there would be no need to adjust the FDR for multiple comparisons across scales. Controlling for the FDR at each scale independently would also guarantee that the FDR would be controlled overall, for all tests performed over all connections and at all scales. This strategy however only makes sense if the objective of the investigator is to systematically examine the results of MSPC at every tested scale. It would also be possible to select the one scale associated with the largest rate of discoveries, defined as the number of significant connections divided by the total number of distinct connections at a given scale. Although this empirical optimization of the scale for means of discovery is attractive in practice, there is however no theoretical guarantee that the FDR at the scale with maximal discovery rate will be controlled. We examined the behaviour of the FDR at “peak of discovery” on simulations, see Section 3.
2.7. Omnibus test

As stated above, control of the overall FDR (across connections and scales) can only be attained with the proposed procedure in the presence of true positives. We thus implemented a permutation procedure on the GLM analysis to perform an omnibus test for the presence of signal across all scales. This test proceeds by comparing the total volume of discoveries across scales observed empirically in the group sample against the total volume of discoveries that could be observed under the global null hypothesis ($G_0$) of no significant discoveries at any connection and any scale associated with a covariate of interest. Formally, for a grid of scales, e.g. 10 to 200 with a step of 10, for a given FDR procedure and for a given admissible level of FDR $\alpha$, let $V_s$ be the volume of discoveries at scale $s$ and $V$ be the total volume of discoveries across all scales. We decided to measure the volume of discoveries at a given scale as the sum of significant squared $t$-tests at this scale, in the event where the FDR pro-

Figure 2: General linear model applied to connectomes at multiple spatial scales. The generation of data-driven brain parcels is iterated at different scales (number of networks), using the bootstrap analysis of stable clustered (BASC), with a hierarchical clustering using the Ward’s criterion. The statistical parametric connectomes are represented using both their real size (left column) and after rescaling to fit identical size (middle column) to illustrate the quadratic increase in the number of connections (multiple comparisons) that comes with an increase in the number of parcels. The results presented here are for illustration purpose only, and not related to the results presented in the application sections of the manuscript.
cEDURE detected discoveries, or the maximal squared $t$-test if no discovery was made. To test the significance of $V$, we generated replication of the statistical parameteric connectomes under the global null hypothesis. For this purpose, the GLM was applied on the original data, then a permutation of the residuals was generated as described in Anderson (2002), see Appendix D. Finally, the replication was generated by adding the permuted residuals to the estimated mixture of explanatory variables, after the explanatory variable of interest (as selected through the contrast) had been removed. In order to respect the dependencies between scales, the same permutation of the subjects was applied to all of the scales for each replication. The detection procedure was applied on each replication and the total volume of discoveries was derived. A Monte-Carlo approximation, with typically 10000 permutation samples, was used to estimate a false-positive rate $p$ when testing against the global null hypothesis. If this test passed significance, then each scale was examined with a control of the FDR at $\alpha = 0.05$, uncorrected for multiple comparisons across scales. If the test did not reach significance, then no connection at any scale was deemed significant.

3. Evaluation on simulated datasets

3.1. Data generation procedure

Rationale. As we were not aware of an existing framework to simulate well-controlled changes in multiscale fMRI network connectivity, we developed one, outlined below. When working at a single scale, it would be straightforward to simulate a change between two populations by changing the values of specific connections in selected individuals, e.g. (Zalesky et al., 2010a). For multiscale connectomes, however, it is important to simulate changes at the level of the fMRI time series, rather than the connectivity values, in order to assess how the simulated changes propagate across scales. In that perspective it is also important to introduce spatial correlations in the simulation: with white noise, the simulated effects get amplified dramatically through averaging on low-scale networks, yet this amplification is much less pronounced in the presence of spatial auto-correlation in the time series. Finally, because manipulating the time series of one network in a simulation can potentially affect many connections in the multiscale connectome architecture, a specific strategy needs to be implemented to precisely control true positives and the effect size at any given scale.

Main procedure. To incorporate some realistic spatial correlations in the simulations, semi-synthetic datasets were generated starting from a large real sample (Cambridge) released as part of the 1000 functional connectome project (Biswal et al., 2010). This sample (Liu et al., 2009) includes resting-state fMRI time series (eyes opened, TR of 3 seconds, 119 volumes per subject) collected with a 3T scanner on 198 healthy subjects (75 males), with an age ranging from 18 to 30

\[\text{http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html}\]
years. All the datasets were preprocessed and resampled in stereotaxic space, as described in Section 4.2. A region growing algorithm was used to extract 483 regions, as described in Bellec et al. (2010). For each subject, a functional connectome was generated (see Section 2.2). The average connectome across all subjects was derived, and a hierarchical clustering procedure (with Ward’s criterion) was applied to derive a hierarchy of resting-state networks at all possible scales, ranging from 1 to 483. The simulation procedure relied on the manual selection of a critical scale $K$ and a particular cluster $k$ at this scale. For each simulation, two non-overlapping subgroups of subject ($N$ subjects per group) were randomly selected. A circular block bootstrap (CBB) procedure was applied to resample the individual time series, using identical time blocks within each cluster, and independent time blocks in different clusters. This resampling scheme ensured that within-cluster correlations were preserved, while between-cluster correlations had a value of zero on average. Finally, for the subjects selected to be in the first group, a single realization of a independent and identically distributed variable, where each time point had a zero mean and a variance of $\sigma^2$, was added to the time series of the regions inside cluster $k$, after the time series were themselves corrected to a zero temporal mean and a variance of $\sqrt{1 - \sigma^2}$. The addition of this signal increased the intra-network connectivity of the cluster including cluster $k$ for all scales smaller or equal to $K$, and increased the within- as well as between-network connectivity for all clusters included in cluster $k$ for scales strictly larger than $k$. Because of the absence of correlations between networks at scale $K$ (due to the CBB resampling), all other connections within- or between clusters at every scale were left unchanged by this procedure. It was thus possible to know exactly which connections were true positive in the group difference at every scale. For each scale, a connection was a true positive (non-zero average difference in connectivity between the two groups) if (1) it connected two subclusters of the cluster of reference; (2) it measured intra-network connectivity for a cluster that was either included in the cluster of reference, or included the cluster of reference. Supplementary Figures S1 and S2 outline the procedure of multiscale connectome simulation.

Effect size and true positive rate. We handpicked two scales (4 and 7), and two clusters at those scales such that the percentage of true positives would be about 15% and 5% respectively. Note that these clusters were used to set true positives at all the scales of analysis, yet the subdivisions (or merging) associated with these clusters represented a varying proportion of the number of clusters at any given scale. As a consequence, the exact percentage of true positives actually varied from scale to scale. Two values for $\sigma^2$ were selected that resulted into a large effect ($\sigma^2 = 0.2$, Cohen’s $d \sim 1$) and a moderate effect ($\sigma^2 = 0.1$, Cohen’s $d \sim 0.5$). The effect size associated with a given $\sigma^2$ actually depended on the within-cluster correlations, between-subject variance in connectivity as well as the scale of analysis. See Supplementary Material Figure S3 for a detailed presentation of the effect size and percentage of true positives as a function of scale.
Simulations under the global null. To assess the behaviour of the testing procedures in the absence of any signal, we also ran experiments under the global null. In that case real connectomes were generated for randomly selected and non-overlapping groups of subjects, and then a testing procedure was implemented to assess the significance of group differences. In these experiments, referred to as the 0% true positive rate experiments, no bootstrap was performed on individual time series nor any signal was added. The experiments simply consisted in comparing real connectomes between random groups of subjects sampled from identical populations.

Robustness to the choice of clusters. Finally, we also investigated how the procedure behaved when the clusters used in the testing procedures did not match exactly with the clusters that were used to generate the simulations. For this purpose, for each simulation, no signal was actually generated in 30% of the regions in the cluster of reference, but rather in an equivalent number of randomly selected regions from other clusters. Although the regions were randomly selected, the same regions were selected across all simulations to simulate a consistent departure of the test clusters from the ground truth clusters. The multiscale clusters without random perturbations were used in the statistical testing procedures. In this setting, many connections outside of the cluster of reference end up with very small effects, and we did not investigate the specificity in this setting given the very large number of true positives and large variations in effect size. However, we did investigate the robustness of the sensitivity of the procedures to perturbation of the clusters, using the same definition of true positives as with the simulations without perturbation.

Simulation scenarios. Our experiments followed a full factorial design with the following parameters:

- moderate ($\alpha^2 = 0.1$) or large ($\alpha^2 = 0.2$) effect size.
- global null (0%), small ($\sim 5\%$) or large ($\sim 15\%$) percentage of true positive.
- $N = 20$ or $N = 50$ subjects per group.
- With/without perturbation of the reference cluster (30 % re-assigned).

For each one of the 16 simulation scenarios, 1000 Monte-Carlo samples were generated and subjected to four statistical detection procedures, described below.

3.2. Methods

Computational environment. In-house tools for all the simulations were implemented using the pipeline system for Octave and Matlab (Belloc et al., 2012).
version 1.0.2, and executed in parallel on the "Guillimin" supercomputer\footnote{http://www.calculquebec.ca/en/resources/compute-servers/guillimin} with CentOS version 6.3 and Octave version 3.8.1. The code can be found on github\footnote{https://github.com/SIMEXP/glm_connectome}.

**Testing procedures.** The MSPC procedure was applied using partitions generated by the group hierarchical clustering algorithm at different scales: from 2 to 10 (step of 1), from 15 to 100 (step of 5), and from 110 to 200 (with a step of 10). For each scale, t-tests and \(p\)-values were generated for the significance of the difference between the two simulated groups at each connection. The FDR procedures were applied with \(q < 0.05\). In addition, a \(p\)-value associated with the global null hypothesis of no effect across all scales was estimated using a permutation procedure with 100 permutation samples. Note that for MSPC, discoveries were deemed significant if the omnibus test passed \(p < 0.05\) and the connection also survived the FDR at \(q < 0.05\). The NBS and MDMR procedures were implemented at a single scale (100 networks). The NBS first used an uncorrected threshold on \(t\)-statistics of 2, followed by a threshold on the size of significant connected components (measured by the extent) with a FWE level of \(p < 0.01\), as proposed in Zalesky et al. (2010a). Note that increase and decrease in connectivity between the two groups were tested independently with this procedure. The MDMR was implemented with a FWE \(p < 0.05\) for the discovery of significant seeds, followed by a local FDR procedure restricted to each connectivity map associated with a significant seed (FDR \(q \leq 0.05\)), as proposed in Shehzad et al. (2014).

**Effective FDR, sensitivity and omnibus test.** The effective FDR at each scale was computed as the proportion of false discoveries divided by the total number of discoveries, averaged across the 1000 replications. The effective sensitivity was computed at each scale as the proportion of discoveries of true positives, divided by the number of true positives present at this scale, average across the 1000 replications. We also examined the distribution of \(p\)-values for the omnibus tests against the global null hypothesis, averaged across the 1000 replications, using regular bins of width 0.05 covering the \([0, 1]\) interval.

3.3. Results

**Empirical false-discovery rate.** Figure\footref{fig:3} represents the empirical FDR as a function of scale for the MSPC procedure. As expected, with the BH FDR, the procedure was slightly conservative, i.e. the empirical FDR was lower than the requested FDR (indicated by a black line). With the group FDR, MSPC was slightly too liberal (indicated by a red line). The group FDR procedure seemed to be more liberal when many discoveries were made, either at the peak of sensitivity across scales, or overall for scenarios with large sample size and effects. Note that the group FDR procedure still had an empirical FDR close or smaller that the nominal value in most configurations, and smaller than 0.1 in all tested configurations.
Figure 3: **Effective false discovery rate (FDR) in simulations, as a function of scale.** The effective FDR of the MSPC was assessed on simulations when testing for group differences. All plotted simulations had a perfect match between the true and test clusters, such that true positive were perfectly characterized. Eight simulation scenarios were investigated: (20 or 50 subjects per group) x (∼5% or ∼15% percentage of true positives) x (moderate or large effect size). The effective FDR was plotted as a function of scale for the BH FDR procedure (blue curve) and the group FDR procedure (red curve). The nominal FDR level set for the procedure was 0.05, as indicated by a flat black line.

**Sensitivity.** Figure 4 presents the sensitivity of the detection as a function of scale (blue: BH FDR, red: group FDR). In all scenarios, the sensitivity was strongly dependent on scale and sharply decreased after reaching a peak at relatively low scales. The position of the peak was dependent on sample size, effect size, and the degree of overlap between the true and the test clusters. When no perturbation of the reference cluster was applied, the group FDR gives superior sensitivity for all but very low scales, were the BH FDR has a sharp peak. In this setting, the sensitivity was excellent for most tested configurations, with sensitivity above or close to 80% as soon as the sample size was large (n = 100), or that the effect size was large. The sharp sensitivity peak for BH FDR strongly decreased when a perturbation was applied on the reference cluster, in which case no substantial advantage was observed for the BH FDR over the group FDR. In the presence of perturbation, the procedure only seemed to reach above 80% sensitivity for n = 100 and large effect sizes, or moderate but highly distributed (50% true positives) effects.

**Testing the global null hypothesis.** Figure 5 presents the distribution (normalized histogram) of the estimated p-values associated with the global null for the different scenarios. For the scenario that actually conforms to the global null (0% true discovery, left-most column), the distribution is more or less uniform, as expected. The actual probability to have p < 0.1 is about 0.06, p < 0.5 is about 0.03 and p < 0.01 is about 0. As soon as there is any large signal, the p-values are systematically smaller than 0.05 in all B = 100 simulations, which means that the overall detection power is outstanding.
The sensitivity of the MSPC was assessed on simulations when testing for group differences. Sixteen simulation scenarios were investigated: (20 or 50 subjects per group) x (≈ 5% or ≈ 15% percentage of true positives) x (moderate or large effect size) x (with or without perfect match between true and test clusters). The true positives used to estimate sensitivity were defined by the reference clusters for the simulation, regardless of the potential perturbation of these clusters prior to simulation to create a mismatch between true and test clusters. The sensitivity was plotted as a function of scale for the BH FDR procedure (blue curve) and the group FDR procedure (red curve).

Comparison between MSPC, NBS and MDMR: specificity. Figure 6 summarizes the empirical false-discovery rate of the different methods investigated here for the different scenarios, but only in the configurations where the true and test clusters exactly match so that true positives are clearly defined. For MSPC, the group and BH FDR procedures behaved very similarly. The empirical FDR across scales was very close to the nominal values. This confirmed empirically that in the presence of signal, controlling for the FDR per scale also controls for the total FDR across scales. When considering the peak of discoveries, the empirical FDR was found to be too liberal yet in reasonable proportion, with empirical FDR up to about 0.1 for a nominal FDR of 0.05. The MDMR was found liberal in situations where there was little signal, but accurate in its control of the FDR otherwise. Note that no cluster-level FWE statistics was applied in MDMR, which could explain the behaviour of the method in low signal configurations. Finally, the NBS method generally resulted in large FDR, up to 0.3. This is not surprising, as the authors of the method emphasized that
null hypothesis ~5% of true positive ~15% of true positive

20 subjects per group (n = 40) 50 subjects per group (n = 100)

"moderate" effect size "large" effect size

<table>
<thead>
<tr>
<th>Group</th>
<th>FDR</th>
<th>BH</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Figure 5: Distribution of p-values for the omnibus test against the global null hypothesis, on simulations. The distribution of p values testing for the global null hypothesis of no association across all scales was assessed on simulations when testing for group differences. Twelve simulation scenarios were investigated: (20 or 50 subjects per group) x (global null, or ~5% or ~15% percentage of true positives) x (moderate or large effect size). Both the BH and group FDR variants of MSPC were applied to each scenario.

NBS only controls for false positives inside a cluster, but not at the voxel level.

Comparison between MSPC, NBS and MDMR: sensitivity. Figure 6 also summarizes the sensitivity of the different methods investigated here for the different scenarios, including with/without perturbations of the reference cluster. The multiscale FDR procedures, either overall across all scales or at the scale with peak percentage of discoveries, had excellent sensitivity profiles. When test and ground truth clusters exactly matched (i.e. no perturbation of the reference cluster), the sensitivity at the peak percentage of discovery was extremely good, even more so with the BH FDR than with the group FDR. This considerable advantage however disappeared when perturbation of the reference cluster was applied. In this configuration, the group FDR dominated for sensitivity across scales, and also in a number of scenarios for the sensitivity at the peak of discoveries. See Tables S1 and S2 for the numerical values represented in Figure 6.

4. Application to real datasets

4.1. Data samples

Participants. The SCHIZO dataset was contributed by the Center for Biomedical Research Excellence (COBRE) to the 1000 functional connectome project

[http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html](http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html)
Figure 6: Summary of specificity (effective FDR) and sensitivity for all testing procedures on simulations. Results from eight simulation scenarios were summarized here: (20 or 50 subjects per group) x (5% or 15% percentage of true positives) x (moderate or large effect size). In addition, the sensitivity was investigated when true and test clusters exactly matched (top right corner) as well as in the more realistic presence of a mismatch between test and true clusters (bottom right corner). Six detection strategies were investigated: the MSPC with the BH and group FDR, either looking just at the scale with peak percentage of discoveries (max) or across all scales (total), the NBS and the MDMR.

The sample comprised 72 patients diagnosed with schizophrenia (58 males, age range = 18-65 yrs) and 74 healthy controls (51 males, age range = 18-65 yrs). The BLIND (Collignon et al., 2011) and MOTOR (Albouy et al., 2012) datasets were acquired at the Functional Neuroimaging Unit, at the Institut Universitaire de Gériatrie de Montréal, Canada. Participants gave their written informed consent to take part in the studies, which were approved by the research ethics board of the Quebec Bio-Imaging Network (BLIND, MOTOR), as well as the ethics board of the Centre for Interdisciplinary Research in Rehabilitation of Greater Montreal (BLIND). The BLIND dataset was composed of 14 congenitally blind volunteers recruited through the Nazareth and Louis Braille Institute (10 males, age range = 26-61 yrs) and 17 sighted controls (8 males, age range = 23-60 yrs). The MOTOR sample included 54 healthy
young participants (33 males, age range = 19-33 yrs).

**Acquisition.** Resting-state fMRI scans were acquired on a 3T Siemens TrioTim for all datasets. One single run was obtained per subject for either the SCHIZO or BLIND dataset while two runs were acquired in each subject for the MOTOR dataset, one immediately preceding and one following the practice on a motor task. For the SCHIZO dataset, 150 EPI blood-oxygen level dependent (BOLD) volumes were obtained in 5 mins (TR = 2 s, TE = 29 ms, FA = 75°, 32 slices, voxel size = 3x3x4 mm³, matrix size = 64x64, FOV = mm²), and a structural image was acquired using a multi-echo MPRAGE sequence (TR = 2.53 s, TE = 1.64/3.5/5.36/7.22/9.08 ms, FA = 7°, 176 slices, voxel size = 1x1x1 mm³, matrix size = 256x256, FOV = 256x256 mm²). For the BLIND dataset, 136 EPI BOLD volumes were acquired in 5 mins (TR = 2.2s, TE = 30 ms, FA = 90°, 35 slices, voxel size = 3x3x3.2 mm³, gap = 25%, matrix size = 64x64, FOV = 192x192 mm²), and a structural image was acquired using a MPRAGE sequence (TR = 2.3 s, TE = 2.91 ms, FA = 9°, 160 slices, voxel size = 1x1x1.2 mm³, matrix size = 240x256, FOV = 256x256 mm²). For the MOTOR dataset, 150 EPI volumes were recorded in 6 mins 40 s (TR = 2.65s, TE = 30ms, FA = 90°, 43 slices, voxel size = 3.4x3.4x3 mm³, gap = 10%, matrix size = 64x64, FOV = 220x220 mm²), and a structural image was acquired using a MPRAGE sequence (TR = 2.3 s, TE = 2.98 ms, FA = 9°, 176 slices, voxel size = 1x1x1 mm³, matrix size = 256x256, FOV = 256x256 mm²).

**Motor task.** Between the two rest runs of the MOTOR experiment, subjects were scanned while performing a motor sequence learning task with their left non-dominant hand. 14 blocks of motor practice were interspersed with 15s rest epochs. Motor blocks required subjects to perform 60 finger movements, ideally corresponding to 12 correct five-element finger sequence. The duration of the practice blocks decreased as learning progressed. It should be noted that the effect of motor learning per se on the subsequent rest run could not be distinguished from that of a mere motor practice/fatigue effect in the present experimental setting.

### 4.2. Methods

**Computational environment.** The datasets were analysed using the NeuroImaging Analysis Kit (NIAK) version 0.12.14, under CentOS version 6.3 with Octave version 3.8.1 and the Minc toolkit version 0.3.18. Analyses were executed in parallel on the “Guillimin” supercomputer using the pipeline system for Octave and Matlab (Bellec et al., 2012), version 1.0.2. The scripts used for processing can be found on github.

---

http://www.nitrc.org/projects/niak/
http://gnu.octave.org/
http://www.bic.mni.mcgill.ca/ServicesSoftware/ServicesSoftwareMincToolKit
https://github.com/SIMEXP/glm_connectome
Preprocessing. Each fMRI dataset was corrected for inter-slice difference in acquisition time and the parameters of a rigid-body motion were estimated for each time frame. Rigid-body motion was estimated within as well as between runs, using the median volume of the first run as a target. The median volume of one selected fMRI run for each subject was coregistered with a T1 individual scan using Minctracc ([Collins and Evans, 1997]), which was itself non-linearly transformed to the Montreal Neurological Institute (MNI) template ([Fonov et al., 2011]) using the CIVET pipeline ([Ad-Dab'bagh et al., 2006]). The MNI symmetric template was generated from the ICBM152 sample of 152 young adults, after 40 iterations of non-linear coregistration. The rigid-body transform, fMRI-to-T1 transform and T1-to-stereotaxic transform were all combined, and the functional volumes were resampled in the MNI space at a 3 mm isotropic resolution. The scrubbing method of [Power et al., 2012], was used to remove the volumes with excessive motion (frame displacement greater than 0.5 mm). A minimum number of 60 unscrubbed volumes per run, corresponding to \( \sim 180 \) s of acquisition, was then required for further analysis. For this reason, some subjects were rejected from the subsequent analyses: 16 controls and 29 schizophrenia patients in the SCHIZO dataset (none in either the BLIND or MOTOR datasets). The following nuisance parameters were regressed out from the time series at each voxel: slow time drifts (basis of discrete cosines with a 0.01 Hz high-pass cut-off), average signals in conservative masks of the white matter and the lateral ventricles as well as the first principal components (95\% energy) of the six rigid-body motion parameters and their squares ([Giove et al., 2009]). The fMRI volumes were finally spatially smoothed with a 6 mm isotropic Gaussian blurring kernel.

Statistical analysis. Brain parcellations were derived using BASC separately for each dataset, while pooling the patient and control groups in the SCHIZO and BLIND datasets, and runs in the MOTOR dataset, over a grid of scales: from 5 to 100 (step of 5), from 110 to 200 (step of 10), and from 220 to 400 (step of 20). By contrast, as the NBS and MDMR procedures were designed to work at a single scale, we implemented them at two scales only, one low and one moderately high: 10 and 100 parcels. The MSPC (with both BH and group FDR), as well as the NBS and MDMR procedures were implemented with the same parameters as those used on simulations (see Section 3.2). For every contrast, the group model included an intercept, the age and sex of participants as well the average frame displacement of the runs involved in the analysis (two covariates of frame displacement were used in the MOTOR dataset, one per run). The contrast of interest was on a dummy covariate coding for the difference in average connectivity between the two groups for SCHIZO and BLIND, and on the intercept (average of the difference in connectivity pre- and post-training) for the MOTOR dataset. All covariates except the intercept were corrected to a zero mean.
Figure 7: **Percentages of discovery as a function of scales and types of FDR correction in the real datasets.** Plots show the rates of discoveries, that is the percentages of connections that show a significant effect in respectively the MOTOR, BLIND and SCHIZO contrasts. Percentages of discovery are shown both for the group and BH FDR, and for all 40 scales. Scales 10 and 100 were selected for further investigation, as highlighted with a gray background.

4.3. Results

**Multiscale discoveries.** The MSPC detection generated maximal percentages of discoveries at low scales (around 10 parcels) for the three datasets, both for the BH and the group FDR. The overall trend was that the rate of discoveries decreased as the number of parcels increased, but with a plateau at low scales, up to 50-100 clusters (see Figure 7). These relationships between discovery rate and scales were similar to the ones observed on simulations, in particular when the test and true clusters did not exactly match (Figure 4). Although the relative changes in discovery rate as a function of scale were similar for the three datasets, the absolute percentages of discoveries were quite different (see Figure 7). The omnibus test was still significant ($p < 0.01$) for all three contrasts, both with BH and group FDR. At scale 10, using the group FDR, the percentages of discoveries were 36%, 5% and 15%, for the SCHIZO, BLIND and MOTOR contrasts, respectively. These rates were almost systematically better than with the BH FDR (30%, 0% and 15%), MDMR (20%, 0%, 10%), and NBS (0%, 0%, 30%). NBS only identified more discoveries when there was an overwhelmingly large discovery rate in the SCHIZO contrast. At scale 100, these percentages fell down to around 21%, 2% and 5% with the group FDR, again almost systematically better than the BH FDR (15%, 1% and 2%), the MDMR (15%, 0.5%, 1%) and the NBS (0%, 0%, 30%). The only exception was again NBS for the SCHIZO sample.

**Spatial distribution of significant discoveries.** Discovery percentage maps revealed which parcels were associated with the largest amount of discoveries, and more specifically the proportion of significant connections for any given parcel,
Figure 8: Percentage of discovery maps in the three real datasets for two scales and four methods. Surface maps show the percentage of connections with a significant effect, for each brain parcel, in respectively the MOTOR, BLIND and SCHIZO contrasts. Percentage discovery maps are shown at two scales (10 and 100) and for four methods (MSPC with group FDR and BH FDR, MDMR and NBS). n.s. = non significant. Maps are projected onto the MNI 2009 surface. See Supplementary Figure S4 for volumetric representations showing results at the subcortical level.

in all three real datasets (See Figure S for surface representations of the cerebral cortex, and Supplementary Figure S4 for volumetric representations including the basal ganglia and cerebellum). Results for scales 10 and 100 were reported for all four explored techniques (MSPC with BH and group FDR, MDMR and NBS). Significant effects were found for all contrasts both at scales 10 and 100 with the group FDR. Widespread effects were observed for the SCHIZO dataset at both cortical and subcortical levels. Parcels with the highest discovery rate were found in the temporal lobe, the prefrontal cortex, the medial temporal lobe and the basal ganglia. The BLIND contrast revealed more localized effects in the occipital cortex, and to a lesser extent in the temporal and premotor cortices. Finally, the MOTOR contrast identified significant effects within an extended cortico-subcortical motor network (sensorimotor and premotor areas, basal ganglia and spinocerebellum). Despite the higher rate of discoveries for the group FDR as compared to the other techniques, the spatial distribution of discoveries were consistent between the four procedures for all contrasts. It is interesting to note that some contrasts that did not show any significant effects using NBS with a \( p < 0.01 \) showed spatially consistent effects when tolerating a very liberal \( p < 0.2 \) (see Supplementary Figure S5). Also, while the discovery rates were larger at scale 10 than scale 100, the spatial distributions were overall consistent between the two scales. Specific parcels nonetheless had higher percentage of discovery at scale 100 than scale 10, e.g. the dorsolateral prefrontal
cortex for the BLIND contrast with the group FDR.

Figure 9: **Group FDR-corrected t-test maps in the SCHIZO dataset for a selection of four seeds at two scales.** T-test maps show significant alterations in functional connectivity (decreases and increases) in schizophrenia for four seed parcels with maximal effects at scales 10 and 100. Maps are projected onto the MNI 2009 surface. See Supplementary Figure S6 for volumetric representations showing results at the subcortical level and the spatial extents of the seeds.

**Seed-based maps of t-statistics.** The maps of discovery rate did not characterize which specific connections were identified as significant, nor the direction of the effect (i.e. an increase vs a decrease in connectivity). We illustrated how these questions can be explored using the SCHIZO dataset, as it showed widespread changes in functional connectivity. The percentage of discovery maps were used to select, both at scales 10 and 100, a number of seed parcels showing maximal effects. These included parcels found in the temporal and prefrontal cortices, the medial temporal lobe as well as the basal ganglia, see Figure 9 and Supplementary Figure S6. For each parcel, a FDR-corrected t-test map associated with the contrast of interest was generated. These t-test maps revealed that the alterations in functional coupling in schizophrenia essentially took the form of a decrease in connectivity, in particular for connections linking temporal, prefrontal and occipital regions. By contrast, the basal ganglia showed an increase in functional connectivity with occipital and temporal areas in schizophrenia. Functional connectivity was also increased between the hippocampus/amygdala and the medial prefrontal cortex. Of note, discoveries made with large networks, at scale 10, were overall consistent with the discoveries made with local regions at scale 100. However, the analysis at high scales sometimes identified stronger, more widespread effects, in specific regions as compared to the analysis at low scales. For example, a strong increase in connectivity between the thalami and
many cortical regions was observed at scale 100, which was smoothed out inside the basal ganglia at scale 10.

5. Discussion

5.1. Multiscale statistical testing

The main contribution of this work has been to examine how the results of a GLM analysis on connectomes were impacted by the number of brain parcels (or scale). We have also proposed a statistically valid procedure for performing such multiscale analyses. We found that the control of the FDR across all tested scales was feasible, and simply followed the control of the FDR within each scale as long as the presence of signal had been established through an omnibus test. On simulations, sensitivity was good within and across scales, allowing investigators to explore the results of the GLM analysis at every selected scale. The sensitivity was excellent when only considering the scale associated with the peak in discovery rate. This latter strategy was however found to be slightly too liberal, and for statistically stringent analysis the peak scale should be selected on an independent dataset, different from the one used in the final analysis. Such departure of the FDR from nominal levels may still be regarded as tolerable for some purposes, e.g. early exploratory studies, as it comes with substantial gains in sensitivity. We would like to emphasize that the objective of the work was to explore multiscale connectomes, rather than identify a single “optimal” scale of analysis. We indeed observed that effects in some brain parcels were much better captured at higher scales, even if the overall rate of discovery was higher for lower scales. We speculate that different scales of parcellation may work better for signals with different biological origins, i.e. synchrony across distributed networks might benefit from a lower scale, while more specialized, local assemblies would benefit from higher scales. Such synchrony of neuronal assemblies can co-exist at different frequency bands within a single spatial location (Varela et al., 2001), and could potentially show distinct, or even opposite, patterns of association with variables of interest. Multiscale GLM analyses offer the only way to detect these different yet complementary aspects of functional assemblies.

5.2. Findings on real datasets

The effects found from analysing the real datasets were consistent with the existing literature. First, schizophrenia has been defined as a dysconnectivity syndrome, with aberrant functional interactions between brain regions being a core feature of this mental illness (for reviews, see Calhoun et al., 2009; Pettersson-Yeo et al., 2011; Fornito et al., 2012). Consistent with previous works, alterations in connectivity were widespread and mostly exhibited decreased connectivity in patients, with the strongest effects in the temporal and prefrontal cortices. Increases in connectivity were also observed, for instance in the basal ganglia, in line with Damaraju et al. (2014). Second, resting-state fMRI studies have previously shown that congenital blindness is associated with
a reorganization of the interactions between the occipital cortex and other parts of the brain, in particular the auditory and premotor cortices (Liu et al., 2007; Qin et al., 2013, 2014), consistent with our findings. Finally, our results are in agreement with the observation that brain activity at rest is modulated by previous intensive motor practice (Albert et al., 2009; Vahdat et al., 2011; Sami et al., 2014). Even in the absence of a definite ground truth on these real life applications, our findings thus had an excellent face validity, and suggested that the MSPC method could be successfully applied to a variety of clinical or experimental conditions.

5.3. Comparison with other methods

With the simulation scenarios we tested, the MSPC method emerged as the most sensitive. Of note is the introduction of the group FDR procedure, which was applied here for the first time to CWAS. This approach proposed by Hu et al. (2010) was designed for improved sensitivity in the presence of sparse signals, and turned out to be beneficial when compared to the widely popular BH FDR procedure (Benjamini and Hochberg, 1995), in all simulation experiments (except at very low scales and unrealistic scenarios). Our evaluation on three real datasets featured different sample sizes: around 15 and 70 participants per group for between-subject comparisons (BLIND and SCHIZO, respectively) and around 50 participants for the within-subject comparison (MOTOR). In addition, while one contrast was expected to involve very distributed systems (SCHIZO), others were anticipated to show fairly localized effects (BLIND and MOTOR). Despite the diversity of those configurations, the comparison between methods on those real datasets led to rather unequivocal conclusions. While the MSPC yielded many significant effects with excellent face validity in all three datasets, MDMR (Shehzad et al., 2014) generated substantially less discoveries. NBS (Zalesky et al., 2010a) only provided significant findings in a few cases (which were found in the presence of very highly significant effects). Overall, the MSPC with group FDR thus appeared to be the most sensitive technique, as compared to the MSPC with BH FDR, NBS or MDMR. In almost all cases, this observation was true both at low (10) and moderately high (100) scales.

As a cautionary note, comparing MSPC with other techniques such as NBS or MDMR is not straightforward because they were not designed to answer the exact same questions, nor do they control for statistical risk in similar ways. In particular, in previous reports, MDMR has been shown to be sensitive at much higher scales (up to voxel resolution) than are tested herein. By contrast, the MSPC behaved particularly well at relatively low scales (tens of parcels), but we expect it to perform poorly at voxel resolution. An interesting future extension of this work would be to integrate MDMR to perform the within-scale testing in our multiscale framework instead of the BH or group FDR procedures. We would first need to establish how statistical risk propagates across scales using MDMR, but our initial findings that MDMR controls FDR well within scale, in the presence of signal, are encouraging.
5.4. Beyond scale selection: choice of the brain parcellation

Although the impact of the number of parcels on CWAS sensitivity has been extensively investigated in this work, we only briefly examined how the choice of parcels, and not just their number, could impact statistical sensitivity. We could, for example, have used random parcellations, like (Zalesky et al., 2010b), or a parcellation based on anatomical landmarks such as the AAL atlas (Tzourio-Mazoyer et al., 2002). From our results on simulations, it seems clear that dramatic differences in statistical power can be achieved at a given spatial scale, if a set of parcels is best adapted to the spatial distribution of an effect. Following an idea initially explored in (Thirion et al., 2006), it may even be possible to relax the constraint of identical parcels across subjects, by matching different individual-specific parcels and use this correspondence to run group-level CWAS analysis. We believe that important improvement in sensitivity could be gained from the optimization of the parcellation scheme, rather than scale, and this represents an important avenue for future research.

5.5. Yet another multiple comparisons problem: multiple contrasts

There is one additional source of multiple comparisons that has not yet been discussed in this paper - the number of contrasts tested in a given experiment. The GLM is very flexible, and by selecting different covariates of interest, the number of contrasts can quickly increase. This problem is not specific to MSPC, but also applies to traditional GLM-based task activation maps. The current consensus seems to treat each contrast as an independent experiment, and to control for multiple comparisons within each contrast. The primary contrast of an experiment can thus be implemented without correction for multiple contrasts, as long as it is clearly defined as part of a data analysis plan, before the results are known. This does not prevent any secondary (exploratory) contrasts, but those additional analyses need to be corrected for multiple contrasts. In our framework, this could be implemented by a Bonferroni correction, which simply divides the significance level of the omnibus tests by the number of contrasts. If a large number of contrasts are tested, however, other, more sensitive approaches may be needed, e.g. Benjamini and Bogomolov (2011).

5.6. Stability and confidence intervals

The MSPC method has been assessed mainly in terms of statistical power. Statistical methods are increasingly being examined as well from the angle of stability, i.e. the ability to replicate the results of an analysis on an independent dataset. High rates of either false positives or false negatives can both be detrimental to stability, but choosing the false-positive rate of a detection in and for itself does not directly translate into a given level of stability. The stability problem has been pervasive and heavily discussed with task-based activation maps, e.g. (Thirion et al., 2007). We did not investigate here how our strategy based on multiscale testing and FDR translated to result stability. The SCHIZO experiment showed that using our approach on a large population sample with a large effect translated into very large rates of discovery. The
right question about stability may not be so much about detecting similar effects across replications, as most connections will be significant, albeit with a small effect size, for large enough datasets, but rather about accurate estimates of the effect size for each connection, i.e. a narrow confidence interval. Future research will incorporate tools for assessing stability into MSPC.

5.7. Other connectivity metrics

Many different metrics have been proposed for measuring functional connectivity between brain parcels, e.g. (Smith et al., 2011). In its current implementation, the MSPC pipeline supports Pearson's correlation, partial correlation (Marrelec et al., 2006), correlation based on a subset of volumes of an individual run, difference in correlation between two subsets of volumes of an individual run, as well as psychophysiological interaction in an individual run. It is also possible to combine multiple runs per subject, e.g. to derive the difference in connectivity between two time points. The MSPC pipeline can be applied to any connectivity metric, provided that (1) it can be estimated on networks of various sizes, and; (2) the distribution of the metric at the group level does not depart markedly from the assumptions of the basic GLM we used: Gaussian, independent and identically distributed (homoscedastic) residuals.

6. Conclusions

In this paper, we introduced the MSPC method for exploring the association between a variable of interest and brain connectivity. The main contribution of the work is to systematically investigate the impact of the number of parcels (spatial scale) on the results of a mass univariate GLM analysis of connectomes, and to propose a feasible way for controlling statistical risk in this context. On simulations, we showed that the proposed technique controlled the FDR appropriately both within and across scales. In addition, the data-driven identification of scales with high rates of discoveries resulted in a better sensitivity than state-of-the-art techniques, such as NBS and MDMR. The MSPC method identified significant results on three real datasets with good face validity - more so than concurrent techniques. Overall, our experiments showed that the MSPC method is both valid and sensitive. The pipeline is available in the NIAK package (Bellec et al., 2011), a free and open-source software that runs in MATLAB and GNU octave. With the emergence of large and phenotypically rich datasets such as NKI enhanced (Nooner et al., 2012) and the human connectome project sample (Van Essen et al., 2013), we believe that the MSPC pipeline will be instrumental in attaining the full potential of these datasets for identifying connectome-phenotype associations.

7. Acknowledgments

Parts of this work were presented at the 2012 and 2013 annual meetings of the organization for human brain mapping, as well as the International Conference
on Resting-State Connectivity 2012 (Magdeburg). The authors are grateful to the members of the 1000 functional connectome consortium for publicly releasing the “Cambridge” and “COBRE” data samples. The computational resources used to perform the data analysis were provided by ComputeCanada[10] and CLUMEQ[11], which is funded in part by NSERC (MRS), FQRNT, and McGill University. This project was funded by NSERC grant number RN000028, a salary award from “Fonds de recherche du Québec – Santé” to PB as well as a salary award by the Canadian Institute of Health Research to PO.

References

URL http://dx.doi.org/10.3389/fnsys.2011.00037


URL http://dx.doi.org/10.1016/j.cub.2009.04.028


URL http://www.worldcat.org/isbn/1852332948


---

[10] https://computecanada.org/
URL http://dx.doi.org/10.3389/fninf.2012.00007

URL http://dx.doi.org/10.1016/j.neuroimage.2010.02.082


URL http://dx.doi.org/10.2307/2674075

URL http://dx.doi.org/10.1073/pnas.0911855107

URL http://dx.doi.org/10.1016/j.neuroimage.2013.03.024

URL http://dx.doi.org/10.1198/jasa.2009.tm08415


URL http://dx.doi.org/10.1109/42.363098

URL http://dx.doi.org/10.1016/j.neuroimage.2011.10.018

Qin, W., Liu, Y., Jiang, T., Yu, C., 2013. The development of visual areas depends differently on visual experience. PloS one 8 (1).


URL http://dx.doi.org/10.1523/jneurosci.4341-13.2014

URL http://dx.doi.org/10.1016/j.neuroimage.2014.02.024

URL http://dx.doi.org/10.1016/j.neuroimage.2009.12.119

URL http://dx.doi.org/10.1016/j.neuroimage.2010.08.063

URL http://dx.doi.org/10.1002/hbm.20210

URL http://dx.doi.org/10.1016/j.neuroimage.2006.11.054

URL http://dx.doi.org/10.1006/nimg.2001.0978


URL http://dx.doi.org/10.1016/j.neuroimage.2013.05.041

URL http://dx.doi.org/10.1038/35067550

URL http://dx.doi.org/10.1002/hbm.20324

URL http://dx.doi.org/10.1006/nimg.1995.1023


URL http://dx.doi.org/10.1371/journal.pone.0015065

URL http://dx.doi.org/10.1016/j.neuroimage.2010.06.041

URL http://dx.doi.org/10.1016/j.neuroimage.2009.12.027

35
Appendix A. Functional connectomes

Let \( \{ P_i, i = 1 \ldots R \} \) be a partition of the brain, i.e. \( R \) parcels such that any voxel in the grey matter belongs to one and only one a parcel. The number of parcels \( R \) is the (spatial) scale of the partition. For an fMRI dataset with \( T \) time samples, the average time series \( w_i \) (vector of length \( T \)) is generated for each parcel \( P \). These average time series are then used to generate a \( R \times R \) matrix of functional connectivity \( Y = (y_{i,j})_{i,j=1}^R \):

\[
y_{ij} = F(\text{corr}(w_i, w_j)), \quad \text{with } F(r) = \frac{1}{2} \log \left( \frac{1-r}{1+r} \right), \quad (A.1)
\]

where \( \text{corr} \) is Pearson’s linear correlation coefficient and \( F \) is the Fisher’s transform. The Fisher’s transform is used to stabilize the variance of the estimated correlation coefficient (Anderson, 1958). This measure of between-network connectivity was used for \( i \neq j \), but we also included a measure of within-cluster average functional connectivity, that uses the voxel-level time series \( w_v \):

\[
y_{ii} = F\left( \frac{1}{\#P_i (\#P_i - 1)} \sum_{v,v' \in P_i, v \neq v'} \text{corr}(w_v, w_{v'}) \right). \quad (A.2)
\]

Appendix B. Ordinary least square GLM estimation

The independent and homoscedastic assumption means that the coefficients of \( E \) are independent from each other and that for each connection \( l \), the \( (e_{n,l})_{n=1}^N \) coefficients are identically distributed with a zero mean and variance \( \sigma_l^2 \). In this context, the maximum likelihood (ordinary least-squares) estimator of \( B \) is:

\[
\hat{B} = (X'X)^{-1}X'Y, \quad (B.1)
\]

and the estimation of the variance of the noise is:

\[
\hat{\sigma}_l^2 = \frac{1}{N-C} \sum_{n=1,\ldots,N} \hat{e}_{n,l}^2, \quad \text{with } \hat{E} = Y - X\hat{B}. \quad (B.2)
\]

For each covariate \( c \), the vector \( (\hat{\beta}_{c,l})_{l=1}^L \) is a vectorized connectome of statistical parameters, quantifying the modulation of each connection \( l \) by the covariate \( c \). This is a direct generalization of the concept of a SPM that has been widely used in task-based fMRI analysis. Each column of the statistical parametric connectome is a actually a SPM at the parcel level (instead of the more standard voxel level), testing the modulation of the functional connectivity of a given seed region with the rest of the brain by the covariate of interest. It is possible to test the significance of each element of the statistical parametric connectome \( \hat{\beta}_{c,l} \) against the null hypothesis (\( H_0 \)) of no association (i.e. \( \beta_{c,l} = 0 \)), using a \( t \)-test:

\[
t_l = \left( \delta_c \hat{B}_l \right) \left( \delta_c^2 (X'X)^{-1} \delta_c \right)^{-1}, \quad (B.3)
\]

36
where \( \delta_c \) is a contrast (column) vector, with \( \delta_d \) equals 1 for \( d = c \) and 0 otherwise. Under \( (H_0) \), the quantity \( t_l \) follows a Student’s \( t \) distribution with \( N - C \) degrees of freedom. By comparing \( t_l \) with the cumulative distribution function \( g_{N-C} \) of the Student’s distribution, it is possible to derive the bilateral probability of observing \( t_l \) under \( (H_0) \):

\[
p_l = 2 \left( 1 - g_{N-C}(|t_l|) \right).
\]

(B.4)

Appendix C. Benjamini-Hochberg and group false-discovery rate

Definition of the FDR. For a given method of selection of significant discoveries, let \( D^F \) be the number of false positive and \( D^T \) the number of true positive. The FDR \( q \) is the mathematical expectation of the ratio between the number of false discoveries and the total number of discoveries \( D^F / (D^F + D^T) \) (with the usual convention that \( 0/0 = 0 \)).

The BH procedure. Let’s first assume that the \( p \) values have been sorted in ascending order, such that \( p_l \leq p_{l+1} \). The BH procedure is built on an estimate \( \hat{q}(p_l) \) of the false-positive rate, equal to \( Lp_l/l \). The \( p_l \) values are screened sequentially, starting with \( l = 1 \), to find the largest \( m \) such that \( \hat{q}(p_l) \leq \alpha \) for all \( l \leq m \). If such an integer does not exist, there are no discoveries. Otherwise, all connections \( l \leq m \) are considered as significant.

The group FDR. The group FDR procedure is based on the separation of the tests \( p \) into families, which where here selected to be seed-based connectivity maps, or in other words the columns of the connectome. The procedure first uses an estimation of the number of true positives per family, and then gives more weight in the test to the family that have the largest estimated rate of discoveries [Hu et al., 2010]. Let us first assume that we knew the proportion \( \pi_0(i) \) of true negatives in the map associated with the parcel \( i \) (oracle case). All the \( p \) values in the map would be re-weighted based on this proportion (in order to give more weight, or smaller \( p \), to maps with lots of signal):

\[
p_{w,i,j} = \frac{\pi_0(i)}{1 - \pi_0(i)} p_{i,j}.
\]

(C.1)

A standard (global) BH procedure is then applied on the re-weighted connectome of \( p \)-values \( (p_{w,i,j})_{i,j=1}^R \), with a modified significance level of \( \alpha/(1 - \pi_0) \), where \( \pi_0 \) is the global proportion of true negatives equal to \( (\sum_{i=1}^R \pi_0(i))/R \). In the event where \( \pi_0(i) \) equals 1, by convention \( p_{w,i,j} \) is not significant for any \( j \).

Estimation of the proportion of signal per map. The group FDR with oracle works under the assumption that the percentage of true negatives is known for each map, while in practice these have to be estimated directly from the data. A number of procedures for estimating the percentage of true negatives have been proposed in the literature, and we selected the LSL estimator in
our experiments (Benjamini and Yekutieli, 2001). For each connection $l$ in the statistical connectome, let $(i,j)$ be the corresponding pair of regions. For a given $i$, let’s assume that the $(p_{ij})$ have been sorted in an increasing order as a function of $j$. The LSL estimator is then given by:

$$
\hat{\pi}_{0,\text{LSL}}(i) = \min_j \left( \left\lfloor \frac{l_{i,j}}{R} \right\rfloor + 1, 1 \right), \quad \text{with} \quad l_{i,j} = \frac{R + 1 - j}{1 - p_{i,j}}. \quad (C.2)
$$

**Enforcing symmetric tests.** Note that, while the initial connectome of $p$-values was symmetric ($p_{i,j} = p_{j,i}$), the weighted matrix of $p_w$-values in the group FDR procedure is not symmetric anymore because the $p_w$-values have been adjusted for the number of potential discoveries in each seed-based connectivity map. It is however not tenable to reach different conclusions about the significant of $p_{i,j}$ and $p_{j,i}$, as these two $p$-values test the significance of identical connections. We thus modified the original group FDR procedure to ensure that, as soon as either $p_{i,j}$ or $p_{j,i}$ reaches significance, both tests were deemed significant. We also adjusted the significance level to $\alpha/2$, where $\alpha$ is the nominal specified FDR, to account for the possible increase in false discoveries due to the symmetrization operation.

**Appendix D. Generation of statistical parametric connectomes under the global null hypothesis**

Let $Y^{(s)}$ be the (subjects x connections) matrix of individual connectomes at scale $s$. Let $\hat{E}^{(s)}$ be the residuals of the regression of $X$ on $Y^{(s)}$, with associated regression coefficients $\hat{B}^{(s)}$. Under the homoscedasticity assumption, the rows (subjects) of $\hat{E}^{(s)}$ are interchangeable, and the resampling scheme first consists of permuting randomly these rows in order to generate a replication $\hat{E}^{(s,\ast)}$ of the residuals. A replication of the connectome matrix under the global null hypothesis ($G_0$) is then generated by recomposing the linear mixture while excluding the $c$-th covariate of interest, tested by the model. Formally, let $X_{\bar{c}}$ and $\hat{B}^{(s)}_{\bar{c}}$ be the reduced model where the $c$-th covariate has been removed from the (subjects x covariates) matrix $X$. Each permutation sample of the dataset is generated as (Anderson, 2002):

$$
Y^{(s,\ast)} = X_{\bar{c}} \hat{B}^{(s)}_{\bar{c}} + \hat{E}^{(s,\ast)}.
$$

(D.1)

The GLM procedure is then implemented with the $Y^{(s,\ast)}$ and the full model $X$ to generate a replication $V^{(s)}_s$ of the volume of discoveries at scale $s$ under ($G_0$).

Because the same dataset at voxel resolution is used to generate all the connectome datasets $(Y^{(s)})_s$, the samples $V^{(s)}_s$ are not independent. In order to respect these dependencies, for any given replication, the same permutation of the subjects is used to generate all of the $(\hat{E}^{(s,\ast)})_s$. The replication of the total volume of discoveries $V^{(s)}$ is then simply the sum of $V^{(s)}_s$ for all $s$. This procedure is repeated $B$ times in order to generate $B$ replications $(V^{(s,\ast)_b})_{b=1}^B$ of the total volume of discoveries under ($G_0$). The Monte-Carlo estimation of the
probability to observe a greater total volume of discoveries under \( (G_0) \) than the actual total volume of discoveries \( V \) generated on the original (non-permuted) dataset is then:

\[
\Pr(V^{(\ast)} \geq V|G_0) \doteq \frac{\# \left\{ b = 1, \ldots, B | V^{(\ast b)} \geq V \right\}}{B}.
\]  

(D.2)

where \( \doteq \) means that the two terms are asymptotically equal as \( B \) tends toward infinity.
Supplementary Material – Multiscale statistical testing for connectome-wide association studies in fMRI

Submitted to Neuroimage.

P. Bellec1,2, Y. Benhajali1,3, F. Carbonell4, C. Dansereau1,2, Z. Shehzad5,6,7, G. Albouy1,8, M. Pelland8, C. Craddock6,7, O. Collignon8, J. Doyon1,8, E. Stip9,10, P. Orban1,9

1Functional Neuroimaging Unit, Centre de Recherche de l’Institut Universitaire de Gériatrie de Montréal
2Department of Computer Science and Operations Research, University of Montreal, Montreal, Quebec, Canada
3Department of Anthropology, University of Montreal, Montreal, Quebec, Canada
4Biospective Incorporated, Montreal, Quebec, Canada
5Department of Psychology, Yale University, New Haven, CT, United States of America
6Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, United States of America
7Center for the Developing Brain, Child Mind Institute, New York, NY, United States of America
8Department of Psychology, University of Montreal, Montreal, Quebec, Canada
9Department of Psychiatry, University of Montreal, Montreal, Quebec, Canada
10Centre Hospitalier de l’Université de Montréal, Montreal, Quebec, Canada

For all questions regarding the paper, please address correspondence to Pierre Bellec, CRIUGM, 4545 Queen Mary, Montreal, QC, H3W 1W5, Canada. Email: pierre.bellec (at) criugm.qc.ca.
Figure S1: **Data generation procedure for simulations.** A hierarchical clustering applied on a group average connectome (n = 198) was used to define partitions at multiple scales. A 7-cluster solution is presented as white squares outlining intra-clusters connections, superimposed on an individual connectome with rows/columns reordered based on the group hierarchy (left panel). A circular block bootstrap scheme is used to resample the original time series. Identical time blocks are used within each cluster, thus preserving intra-cluster connectivity. Independent time blocks are used between clusters, thus setting inter-network connectivity to zero (middle panel). A single simulated time series is added to all the regions belonging to one selected cluster, thus increasing the intra-network connectivity (right panel).

Table S1: **Summary of the effective false-discovery rate of six detection methods on simulations.** The effective FDR of the MSPC and other methods was assessed on simulations when testing for differences in connectivity between groups. All simulations included in the table had a perfect match between the true and test clusters, such that true positive were perfectly characterized. Eight simulation scenarios were investigated: (20 or 50 subjects per group) x (~ 5% or ~ 15% percentage of true positives) x (moderate or large effect size). For each scenario, six detection strategies were investigated: the MSPC with the BH and group FDR, either looking just at the scale with peak percentage of discoveries or across all scales, the NBS and the MDMR.
Figure S2: **Impact of simulated changes on multiscale connectomes.** The left column presents an individual connectome, after circular block resampling, at multiple scales (4, 7, 30, 100). The middle column shows the same connectome after a signal was injected in all regions belonging to one of the clusters at scale 7. The right column is the difference between the middle and the left column. Note how the main and only significant differences are concentrated in the connections that linked clusters that are either subclusters of the cluster of reference, or include the cluster of reference, as outlined by a white square.
Figure S3: **Percentage of true positive and effect size as a function of scale in the various simulation scenarios.** The top row shows the percentage of true positives in the simulation as a function of scale, when the reference cluster was selected at scale 7 (left) or scale 4 (right). The middle and bottom rows show the average effect size over all true positives, for a “moderate” (middle row) and “large” (bottom row) effect sizes. The effect size is plotted for simulations where the test and ground truth clusters exactly matched (blue curve) as well as for simulations where a perturbation of the reference cluster was applied (red curve).
Figure S4: **Percentage of discovery maps in the three real datasets for two scales and four methods, in volumetric space.** Percentage discovery maps show the percentage of connections with a significant effect, for each brain parcel, in respectively the MOTOR, BLIND and SCHIZO contrasts. Maps are shown at two scales (10 and 100) and for four methods (MSPC with group FDR and BH FDR, MDMR and NBS). n.s. = non significant. MNI coordinates are given for representative slices superimposed onto the MNI 152 non-linear template.
Figure S5: Percentage of discovery maps with NBS at a very liberal threshold $p < 0.2$ in the BLIND dataset. Percentage discovery maps show the percentage of connections with a “significant” effect using the NBS method, for each brain parcel, in the BLIND contrast. Maps are shown at two scales (10 and 100). MNI coordinates are given for representative slices superimposed onto the MNI 152 non-linear template.

Figure S6: Group FDR-corrected t-test maps in the SCHIZO dataset for a selection of four seeds at two scales (10 and 100), in volumetric space. The upper row shows the seed parcels selected at scales 10 (large networks) and 100 (local regions). The t-test maps show significant ($q < 0.05$) alterations in functional connectivity (decreases and increases) in schizophrenia for each seed at scales 10 and 100. MNI coordinates are given for representative slices superimposed onto the MNI 152 non-linear template.
Table S2: Summary of the sensitivity of MSPC of six detection methods on simulations. The sensitivity of the MSPC and other methods was assessed on simulations when testing for differences in connectivity between groups. Eight simulation scenarios were investigated: (20 or 50 subjects per group) x (∼5% or ∼15% percentage of true positives) x (moderate or large effect size). In addition, simulations had either a perfect match between the true and test clusters, or a perturbation was applied on the reference cluster in order to create a mismatch. For each scenario, six detection strategies were investigated: the MSPC with the BH and group FDR, either looking just at the scale with peak percentage of discoveries or across all scales, the NBS and the MDMR.