Assessing agreement between experiments to distinguish conditions: Reproducibility of discriminating gene networks across brain tissues in GTEX

A Caceres and JR Gonzalez

March 27, 2017

Abstract

Reproducibility is a fundamental tenet of science yet perceived feeble in current biomedical research. Great effort is put into assessing the concordance between two experiments measured on the same set of individuals under controlled conditions, such as time points, exposures or tissues. However, many studies are designed to measure a population sample under a range of different conditions and, like any scientific study, their results are expected to be reproducible in other samples under different experimental setups. Surprisingly, there is a lack of statistical measures that assess the degree of agreement between independent studies to distinguish conditions. Take for instance the GTEX project, where gene expression data is measured in several tissues using RNA-seq, and validated with expression microarrays on the same individuals. In this study, one can infer the correlation gene network, a fundamental biological entity, for each tissue. It is therefore expected a level of reproducibility to discriminate gene networks between tissues, as derived in other studies of different individuals and experimental procedures. We propose an agreement measure of condition discrimination that generalizes Cohen's kappa, in which the elements of a cross-tabulated table between conditions are pair-wise correlations between two different studies. We derive the distributional characteristics of the measure, and show how it increases monotonically with kappa while its variance allows high precision estimates of intermediate agreement. We use the measure to test the agreement to distinguish between the gene networks of four brain regions as inferred from the GTEX (RNA-seq) and BRAINEAC (microarray) projects. We find full agreement to distinguish between gene networks across tissues and fair agreement for their gene ontology enrichment status. As a conclusion, GTEX unprecedented expression data should be currently used as a benchmark to reproduce tissue specificity of gene networks obtained in independent studies.

1 Introduction

Reproducibility is a pressing issue in biomedical research that particularly worries a large number of researchers in the field [1]. Research guidelines from leading journals and the American Statistician Association urge for the need of confirmation studies and accurate statistical reporting (www.nih.gov/about/reporting-preclinical-research.htm)[2, 3]. In systems biology, interaction networks are often derived from the integration of high-throughput data. A number of metrics exist to test the preservation of the networks under different conditions [4]. If the conditions are different experiments then the measures can be used to assess the reproducibility of the network. However, as in many experiments that aim to test a set of individuals under varying conditions, there is interest to test, for instance, how the interactions of a gene network change between tissues. Clearly, the validity of such experiments also need to be assessed. In these cases, preservation metrics are measurements within experiments to distinguish between conditions while reliability should assess the degree of agreement between experiments to distinguish conditions.

In statistics, there are numerous ways to measure the reliability of an observation. Reliable observations are reproducible and accurate. Agreement measures between two experiments on the same individuals

are used to assess the consistency of the observations being made. If observations are classifications of individuals into groups, Coehn's κ and its generalizations are typically used [5, 6]; if observations are continuous then a number of correlation measures can be used, such as Pearson's or intra-class correlations [7]. These and other similar agreement measures are suitable when experiments are performed under controlled conditions on the individuals. When experiments are designed to test the individuals under a range of varying controlled conditions, it is of interest to test first the ability of the observations to distinguish between conditions and second the consistency of such ability, when experimental procedures and individual samples change. Remarkably, for this type experiments, there is a lack of reliability measures that, in particular, can help us assess the reliability of co-expression gene networks to distinguish between a range of tissues. We therefore propose a generalization of Cohen's κ that measures the agreement between experiments to distinguish conditions.

The GTEX project is an unprecedented effort to measure the gene expression in tens of tissues in hundreds of subjects [8]. It is therefore a strong candidate for becoming a preferred benchmark for the interaction networks inferred in specific experiments. Currently, the validity of a gene or protein network derived from high throughput data is typically assessed by comparing it with networks derived from current knowledge of specific interactions, given by curated pathways, specific experiments, or even text mining of published articles, etc [9]. This type of confirmatory analysis does not take into account, for instance, that some networks may arise in one tissue and not in another. Therefore, while validity is thus investigated, reproducibility is not being measured. Reproducible networks are observables of reproducible experiments, one of which could be taken as GTEX. Agreement measures with analyses derived from this project may then serve as confirmatory experiments for reliability assessments of interactions network [3].

Studies that measure gene expression in a range of tissues can become more common, one of which is the BRAINEAC project [10]. Here, the gene expression using a microarray data was measured in hundreds on un-demented individuals at the time of death in nine different tissues. Using our agreement measure, we therefore investigated the reliability of discriminating gene networks across common brain tissues between BRAINEAC and in GTEX studies. We tested the reliability of genome-wide and KEGG networks (www.genome.jp/kegg).

Given that an important part of reproducibility research is the detailed reporting of the analysis and results [11], all necessary data and the code needed to reproduce all the reported results of this work has been made publicly available in /github/.

2 Methods

We propose an agreement measure of experiments to distinguish between controlled conditions. While the measure can be applied on different research settings, we illustrate how its need arises from an example in current functional genomic research.

2.1 The problem

Let us assume that we have two experiments that measure genome-wide gene expression in two different population samples, in the same range of tissues (conditions). The experimental conditions and setups may also vary, i.e. one experiment uses RNA-seq and the other microarray technologies. We are interested in inferring the co-expression gene networks across tissues and determine whether the networks are consistent between experiments, where co-expression between two genes is determined by their correlation over the subjects' gene expression levels. Table 1 shows the experimental design, where we assume thee genes (variables) measures under three tissues (conditions) in two different population samples.

The interaction network for each of the experiments and conditions can be represented by a correlation matrix between all gene-pairs. Given that the correlation matrix is symmetrical, the network is fully determined by the upper triangular terms of the matrix. Table 2 shows the network derived for a single condition (tissue A) in both experiments.

Experiment 1 - E1

	A				B				C		
	v1	**v2**	**v3**		**v1**	**v2**	**v3**		**v1**	**v2**	**v3**
i1	a11	a12	a13	**i1**	b11	b12	b13	**i1**	c11	c12	c13
i2	a21	a22	a22	**i2**	b21	b22	b23	**i2**	c21	c22	c23
.	.	.	.	.	.	.	.	.	.	.	.
im	am1	am2	am3	**im**	bm1	bm3	bm2	**im**	cm1	cm2	cm3

Experiment 2 - E2

	A				B				C		
	v1	**v2**	**v3**		**v1**	**v2**	**v3**		**v1**	**v2**	**v3**
i1	a'11	a'12	a'13	**i1**	b'11	b'12	b'13	**i1**	c'11	c'12	c'13
i2	a'21	a'22	a'22	**i2**	b'21	b'22	b'23	**i2**	c'21	c'22	c'23
.	.	.	.	.	.	.	.	.	.	.	.
it	a't1	a't2	a't3	**it**	b't1	b't3	b't2	**in**	c't1	c't2	c't3

Table 1: Two experimental measurements ($E1$, $E2$) of 3 variables ($(v1, v2, v3)$, $k = 3$) under three conditions/states ($i = (A, B, C)$, $n = 3$) on m items/individuals

	A				A'			YA	E1	E2
	v1	**v2**	**v3**		**v1**	**v2**	**v3**	**Y1**	A1	A'1
v1	1	A1	A2	**v1**	1	A'1	A'2	**Y2**	A2	A'2
v2	A1	1	A3	**v2**	A'1	1	A'3	.	.	.
v3	A2	A3	1	**v3**	A'2	A'3	1	**Yl**	Al	A'1

Table 2: Network for state A inferred in Experiment 1 (A) and Experiment 2 (A') from correlations between variables shown in matrix form (above) or observation list of l items (correlation pairs between experiments), where $l = \frac{1}{2}(k^2 - k)$.

A measure of the module preservation of the network is the correlation between the triangular terms of the network inferred in each experiment [4], other preservation measures are also possible. From table 2, we therefore compute the correlation between inferred networks

$$c(A, A') = cor(E1(YA), E2(YA)). \tag{1}$$

To assess agreement of networks between experiments we then form the cross-tabulated table of networks between experiments:

	A'	**B'**	**C'**
A	c(A,A')	c(A,B')	c(A,C')
B	c(B,A')	c(B,B')	c(B,C')
C	c(C,A')	c(C,B')	c(C,C')

Table 3: Cross-tabulation of Network correlations between experiments

We would then like to have a measure of the agreement of the cross-tabulated Table 3, whose elements are point estimates with given distributional properties.

2.2 A solution

Cross tabulation for two judges observing m items in n categories takes a similar of Table 3, see Table 4. where $N(X, Y)$ is the number of items measured in category X and Y by the first and second judge respectively. Agreement is typically measured by Cohen's kappa

	A	**B**	**C**
A	N(A,A)	N(A,B)	N(A,C)
B	N(B,A)	N(B,B)	N(B,C)
C	N(C,A)	N(C,B)	N(C,C),

Table 4: Cross-tabulation of measurement scores in categories (A, B, C) on the same group of items between two judges/experiments

$$\kappa = \frac{\sum_{i=1}^{n} P(X_i, X_i) - \sum_{i=1}^{n} P_1(X_i)P_2(X_i)}{1 - \sum_{i=1}^{n} P_1(X_i)P_2(X_i)} \tag{2}$$

where $P(X_i, X_i) = N(X_i, X_i)/n$ is the observed frequency of items that where measured in category X_i by both judges and $P_j(X_i)$ is the frequency of items in X_i observed by judge j ($j = 1, 2$). κ measures the the fraction of discordant observations expected by chance that are actually observed in agreement. The sum $P_0 = \sum_i P(X_i, X_i)$ is the total fraction of agreement, that falls in the diagonal, which does not account for random agreement.

From Table 3, we propose to measure the probability that the diagonal items on the table are their row and column maxima:

- $p_{AA} = Pr(c(A, A') > c(A, B'), c(A, C'), c(B, A'), c(C, A'))$,
- $p_{BB} = Pr(c(B, B') > c(B, A'), c(B, C'), c(A, B'), c(C, B'))$ and
- $p_{CC} = Pr(c(C, C') > c(C, A'), c(C, B'), c(A, C'), c(B, C'))$,

where p_{ii} ($i = A, B, C$) is the probability that the correlation of network i between experiment 1 and Experiment 2 is the maximum of the correlations between the network i in one experiment and any other network in the other experiment. These probabilities can be computed as the product of the individual pair-wise probabilities

$$p_{ii} = \prod_j Pr(c(i, i') > c(i, j')) * Pr(c(i, i') > c(j, i')), \tag{3}$$

Where the first factor is the maximum over rows (Experiment 1), the second factor the maximum over columns (Experiment 2), and the product runs over all other possible conditions (j). If we assume that the correlations $c(a, b')$ can be transformed to normal random variables $z_{ab'}$ using, for example, a Fisher's z transformation, then the probability that the diagonal term (i, i') is higher than other term j' in the row can be computed from

$$Pr(c(i, i') > c(i, j')) = \frac{1}{2}(1 - erf(\frac{1}{\sqrt{2}} \frac{\mu_{ij} - \mu_{ii}}{\sqrt{\sigma_{ii}^2 + \sigma_{ij}^2}}), \tag{4}$$

where erf is the error function. The expression follows from assuming a transformation T such that

$$z_{ij'} = T(c(i, j')) \tag{5}$$

$$z_{ij'} \sim N(\mu_{ij}, \sigma_{ij}^2) \tag{6}$$

and performing the integration over the joint distribution

$$Pr(c(i, i') > c(i, j')) = \int_{-\infty}^{\infty} \int_{z_{ij'}}^{\infty} N(\mu_{ii}, \sigma_{ii}^2) N(\mu_{ij}, \sigma_{ij}^2) dz_{ii'} dz_{ij'}. \tag{7}$$

Therefore, we have that the probability that the diagonal term $c(i,i')$ is the maximum in the row i is

$$\prod_j Pr(c(i,i') > c(i,j')) = \frac{1}{2}\prod_j (1 - erf(\frac{1}{\sqrt{2}}\frac{\mu_{ij} - \mu_{ii}}{\sqrt{\sigma_{ii}^2 + \sigma_{ij}^2}}). \tag{8}$$

The the probability that the diagonal term $c(i,i')$ is the maximum in the column i follows a similar form.

Our agreement measure then follows from the overall probability that the diagonal items on the cross-tabulated table are their row and column maxima. This is the probability of n successes in n Bernoulli trials each of which has its own probability p_{ii}, or a bionomial Poisson distribution with mean and variance

$$\mu = \sum_i p_{ii} \tag{9}$$

$$\sigma^2 = \sum_i p_{ii}(1 - p_{ii}) \tag{10}$$

We define the fraction of successes $\lambda = \mu/n$ with corresponding variance $\sigma_\lambda^2 = \sigma^2/n^2$ as the agreement measure of experiments to distinguish between conditions. In the case that Experiment 1 (in rows) is the benchmark for Experiment 2 (in columns), then one is interested in testing whether the diagonal terms are the maxima of their rows only, generalizing the concepts of sensitivity and specificity for more than two conditions. In this case λ can be computed by simply setting $Pr(c(i,i') > c(j,i') = 1$.

2.3 Comparison between measures of agreement

We compared the performance of the state agreement's λ with κ under different scenarios. We first noted that cross-tabulation in Table 4 for κ measurements can be casted into the cross-tabulated table of inferences as Table 3. Given that for row i the number of observed items is $N_i = P_1(X_i)n$, we can then assume that $N(X_i, X_j)$ is one draw of a binomial variable

$$N(X_i, X_j) \sim Binomial(N(X_i, X_j), N(X_i, X_j)/N_i) \tag{11}$$

with mean, and variance of the mean,

$$\mu_{ij} = N(X_i, X_j) \tag{12}$$

$$\sigma_{ij}^2 = N(X_i, X_j)(1 - N(X_i, X_j)/N_i), \tag{13}$$

which distributes normally for large N_i. These values can be used in equation 8. With a similar computation for the column elements, the measure λ can be obtained for a table in the form of Table 4. The state agreement λ can thus be compared with the value of κ for varying values of the total fraction of agreement P_0. We also compared the measure with the fraction of times a diagonal element is a row and column maximum r, obtained from:

$$R_i = \begin{cases} 1, & \text{if } N(X_i, X_i) = max(\{N(X_i, X_j), N(X_j, X_i)\}_j) \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$r = \frac{1}{n}\sum_i R_i. \tag{15}$$

We performed a series of simulations to compare these four measurements. We selected six different scenarios given by three possible number of states/categories $n = (5, 10, 15)$, and three possible initial forms for the marginal frequencies $P1(X_j)$ and $P2(X_j)$

- Senario 1 (equiprobable): $P1(X_j) = P2(X_j) = \frac{1}{n}$

- Senario 2: $P1(X_j) = P2(X_j) = \frac{1}{j} \sum_j \frac{1}{j}$
- Senario 3 (least equiprobable): $P1(X_j) = P2(X_j) = \frac{1}{j^2} \sum_j \frac{1}{j^2}$

We set the number of observations to $m = 500$. For each scenario, we simulated 50 cases of perfect agreement tables, i.e. diagonal matrices, and 50 cases of perfect disagreement; those are tables with zeros on the diagonal terms except for the cell of maximum joint probability. For each case, we permuted 20 pairs of observations 100 times. After each 20 pairs of permutations, we computed the four agreement measures. This procedure allowed assessment of 5,000 simulations with decreasing agreement from 1 to 0 and 5,000 simulations with increasing agreement from 0 to 1, covering the whole agreement interval.

We used `R.3.30` and the package `psych` to perform calculations and compute the Cohen's κ.

2.4 Gene expression data

We downloaded expression data from the GTEX project obtained from RNA-seq (http://www.gtexportal.org). Reads per kilobase per million mapped reads (RPKM) of version 6 were obtained for all brain tissues (`GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz`). Covariates for each tissue were also downloaded (`GTEx_Analysis_V6_eQTLInputFiles_covariates.tar.gz`).

We also downloaded the brain expression data of the BRAINEAC project (http://www.braineac.org/) obtained from winsorized values of exon array data (Affymetrix Human Exon 1.0 ST array). Downloaded data has been previously normalized and corrected for batch effects.

We identified four brain tissues common in both datasets and for which GTEX had covariate information. Those were cerebellum (CRBL) with 125 individuals in GTEX and 130 in BRAINEAC, frontal cortex (FCTX) with 108 and 135 individuals, (HIPP) hippocampus with 94 and 130 individuals, and putamen (PUTM) with 82 and 135 individuals, respectively. Between the two studies, we mapped 10,683 genes for wich we computed the all the pair-wise correlations between the expression values. We used a partial correlation for the GTEX data, in which we adjusted for the covariates, and a Pearson's correlation for the gene co-expressions in BRAINEAC. Scripts with co-expression analysis code and the required data can be found in .

3 Results

3.1 Simulations

We observed that the condition agreement measure λ increased monotonically with κ for all the simulation scenarios, see figure 1. The functional dependence was highly stable under different scenarios, revealing, as expected, high λ agreement for fair values between $(0.2, 0.4)$ of κ, as the latter is a measure of exact agreement rather than discriminative agreement. For low values, λ tends to zero when κ can take small negative values, a situation already described in Cohen's. We also observed that for a given κ there is a sizable range of λ values, in particular as conditions become less equiprobable (changes of scenarios from 1 to 3). Note that, when the number of conditions is small (5) and the marginal distribution greatly concentrates around a single condition ($j = 5$ for scenario 3) then λ tends to $1/\#conditions$ (0.2), as the experiments can clearly distinguish this condition from the rest. In this case, κ tends to zero.

In terms of λ's variance (figure 2), we found that it decreases with the number of states, and departure from marginal equiprobability. For a given λ, we observed that a range of variances are allowed; whereas κ has a one to one correspondence between mean and variance (not shown). In particular, λ's variance seem to decrease towards zero when the mean of λ tends to r, that is, when the probabilities of diagonal terms of being row maxima tend either to zero or to one. From a practical point of view, this means that if the elements of the cross-tabulated table of inferences (Table 3) are determined each with high precision (low variance) then the agreement measure can also be estimated with high precision, as well. The effect is clearly visible in the scenarios 2 and 3 and low number of conditions. As the number of conditions increase, the effect should be visible with a substantial increase on the number of simulations. When concentration around a single condition is present, we observed a clear reduction of the possible values for the variance, around $\lambda = 1/\#conditions$.

The comparisons between three agreement measures (κ, λ and r) are shown in figure 3 for changing number of conditions for the equiprobable scenario 1, as a function of the accuracy value P_0. We confirmed the lower estimate of κ with respect to P_0 and observed that the difference decreases as the number of states increases. This is as expected since agreement test with lower number of categories are more prone to correct classification due to chance. Similarly r, the fraction of times the diagonal terms are row and column maxima, is higher than λ, a distributional estimate of such fraction.

3.2 Real data

3.3 Genome-wide networks

We inferred the genome-wide co-expression networks between for 10,683 genes across the GTEX and BRAINEAC studies in four brain tissues: cerebellum, frontal cortex, hippocampus and putamen. Each networks was fully characterized by 5.7×10^7 interactions which correspond to the upper triangular matrix terms of the correlation matrix of expression levels. We assessed the agreement between studies to distinguish the genome-wide networks across all four tissues. Figure 4 illustrates the correlations between the networks, whose values where previously z-transformed. We observed that the all correlations were similar in size between (0.37, 0.46). However, their standard errors where small ($\sim 10^{-5}$), given the large number of degrees of freedom. More specifically, the figure shows that the cerebellum and frontal cortex diagonals are the maxima of their rows and columns, and therefore the two studies can discriminate between them. For the hippocampus and putamen, note that they are the second maxima after the correlation of GTEX functional cortex with each of these tissues in BRAINEAC. The experiments then cannon clearly agree on how distinguishable is the frontal cortex from the hippocampus and putamen.

We computed the agreement measure λ from Tables 5 and 6, which are the normally distributed variables, inferred from the between network correlations. As expected form the observations made in figure 4, we obtained a value of $\lambda = 0.5$ and vanishing variance. This value of λ reports that a fraction of $1/2$ conditions are agreed to be different between experiments. The high precision of the estimate follows from the small standard errors of the correlations, due to the large number of degrees of freedom.

We also benchmarked BRAINEAC networks with respect to GTEX. We hence looked only at the diagonal terms within their rows. In this case, we confirmed that all diagonal terms were their row maxima (see table 7), and therefore $\lambda = 1$. These result show that, leaving other confirmatory studies to assess GTEX as a possible benchmark, BRAINEAC fully agrees with GTEX in terms of sensitivity and specificity.

3.4 KEGG networks

The Kyoto encyclopedia of genes and genomes (KEGG) offers a list of experimentally characterized biochemichal pathways. We selected the annotated genes in each study, for the proteins of 292 pathways. For each of these pathways, or subset of genes, we computed the full agreement meassure λ and its benchmark version, similarly to the previous section.

The full agreement λ is shown in table 8 for pathways with the top values ($\lambda > 0.5$). We observe 8 pathways (2%) with agreement between (0.5, 0.75), those are pathways for which there is agreement to be distinguishable in between two and three tissues. No pathways is likely to be different in all four tissues across studies. Interestingly, 5 of these pathways are directly linked with signaling processes specific to brain. The top hit with $\lambda = 0.68$ and $\sigma^2_\lambda = 0.012$, *neuroactive ligand receptor interaction*, is illustrated in figure 5. The figure shows that the cerebellum is not clearly identified by BRAINEAC, as the diagonal term is the minimum in the row. However, a clear distinction is obtained for the frontal cortex, hippocampus and putamen areas with estimate for λ lower than $r = 0.75$ that accounts for sizable uncertainty in the estimates of the correlations.

We also banchmarked BRAINEAC with respect to GTEX for the KEGG pathways. We confirmed the higher estimated of λ in this case, since lower comparisons for the diagonal terms are included, and therefore their probabilities of being row maxima increase. In particular, we observed that 5 pathways had agreement between (0.75, 1), meaning that BRAINEAC can agree to distinguish between 3 to 4 tissues in these pathways, if GTEX variability is not taken into account (Table 9. Three of the four pathways are specific to brain and were previously obtained in the full agreement measure. In particular, *neuroactive*

ligand receptor interaction, increased to $\lambda = 0.805$. We interpret this result as a gained distinction between the frontal cortex, hippocampus and putamen, and an increment in the uncertainty that the diagonal term of the cerebellum is not the row maxima; respect to the full agreement measure.

4 Discussion

We propose a new measure, λ, of agreement between studies. The motivation of the measure is the assessment of agreement between studies on items (subjects, co-expression pairs, etc) under a range of controlled conditions. In particular, we studied the agreement of studies to distinguish between the co-expression networks of four different brain tissues. We are unaware of similar measures of agreement, in particular for testing large interaction networks. Measures of module network preservation allow the assessment for the reliability of one network over different conditions, the correlation between co-expression networks being one of them [4]. While other preservation measures can be used for network discrimination between conditions, here, we are interested in assessing the overall reproducibility of the discrimination in two different experiments. As the new measure is conceptually closer to inter-rater agreement measures, we designed a simulation framework in the properties of λ could be compared with those of Cohen's *kappa*. λ is a suitable reliability measure as it satisfies three basic requirements: i) its values range from 0, null agreement, to 1, perfect agreement, ii) tends to zero when expected agreement tends to zero and iii) it accounts for random agreement. As compared with *kappa*, λ systematically leads to higher agreement. Perfect agreement for κ is exclusively given by diagonal tables, while perfect agreement for λ is given by maxima diagonal terms that are estimated with low variance. This is an important difference between the measures, which allow λ to be utilized in more general situations where the elements of the cross-tabulated table are inferences, and not only the proportion of times two raters agree on a measurement of a set of items. In particular, we observe that λ can be estimated with low variance for intermediate values of agreement, or intermediate fraction of conditions that are distinguishable between studies. Therefore, as λ can be less conservative measure, it allows for a suitable generalization to studies that deal with numerous controlled conditions, that cannot be covered by κ.

In our application to co-expression networks in brain, we found that GTEX and BRAINEAC agree on the discrimination of 2 tissues out of 4, at a genome-wide level. Note that the two studies are based on very different technologies (RNA-seq and microarray) and analysis methods to infer the networks in two different sets of subjects. Our results are unlike other studies, in which these two technologies have been applied within the same study and subject sample to assess the level of agreement on gene expression measurements. While those studies measure the necessary reliability between technologies, our study assesses the reproducibility of between inferences from independent studies. As such, our method is of meta-analysis nature, where consistency between inferences are studied.

We made two further observations. If GTEX is considered as a benchmark study the agreement measures increase. In this case, we assume that GTEX validity as benchmark for gene-network inferences should be considered in other studies. Therefore the study's variability is not considered in the agreement assessment. We also observed that specific biochemical pathways can also be assessed for agreement. This focused approach lead to the identification of pathways specific to the biology of the brain. Our results suggest that agreement assessment can be thus put to the service of finding new biological effects.

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[2] Ronald L Wasserstein and Nicole A Lazar. The asa's statement on p-values: context, process, and purpose. *Am Stat*, 70(2):129–133, 2016.

[3] Jeffrey S Mogil and Malcolm R Macleod. No publication without confirmation. *Nature*, 542(7642):409–411, 2017.

[4] Peter Langfelder, Rui Luo, Michael C Oldham, and Steve Horvath. Is my network module preserved and reproducible? *PLoS Comput Biol*, 7(1):e1001057, 2011.

[5] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[6] Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23, 1999.

[7] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[8] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[9] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.

[10] Daniah Trabzuni, Mina Ryten, Robert Walker, Colin Smith, Sabaena Imran, Adaikalavan Ramasamy, Michael E Weale, and John Hardy. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *Journal of neurochemistry*, 119(2):275–282, 2011.

[11] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Comput Biol*, 9(10):e1003285, 2013.

5 Figures

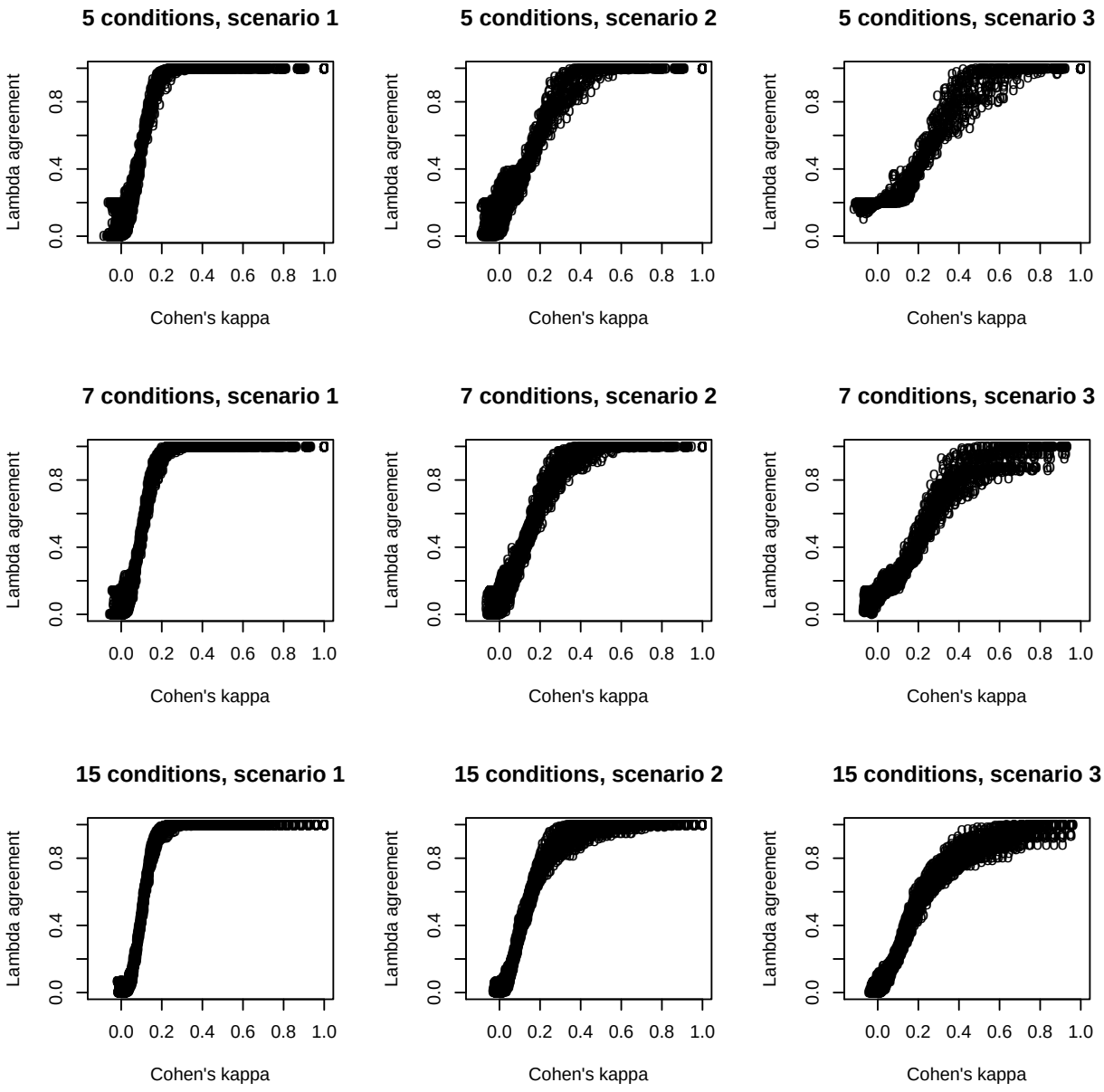

Figure 1: Lambda compared with Cohen's kappa for values of acurracy P0, or total agreeement fraction, ranging from 0 to 1.

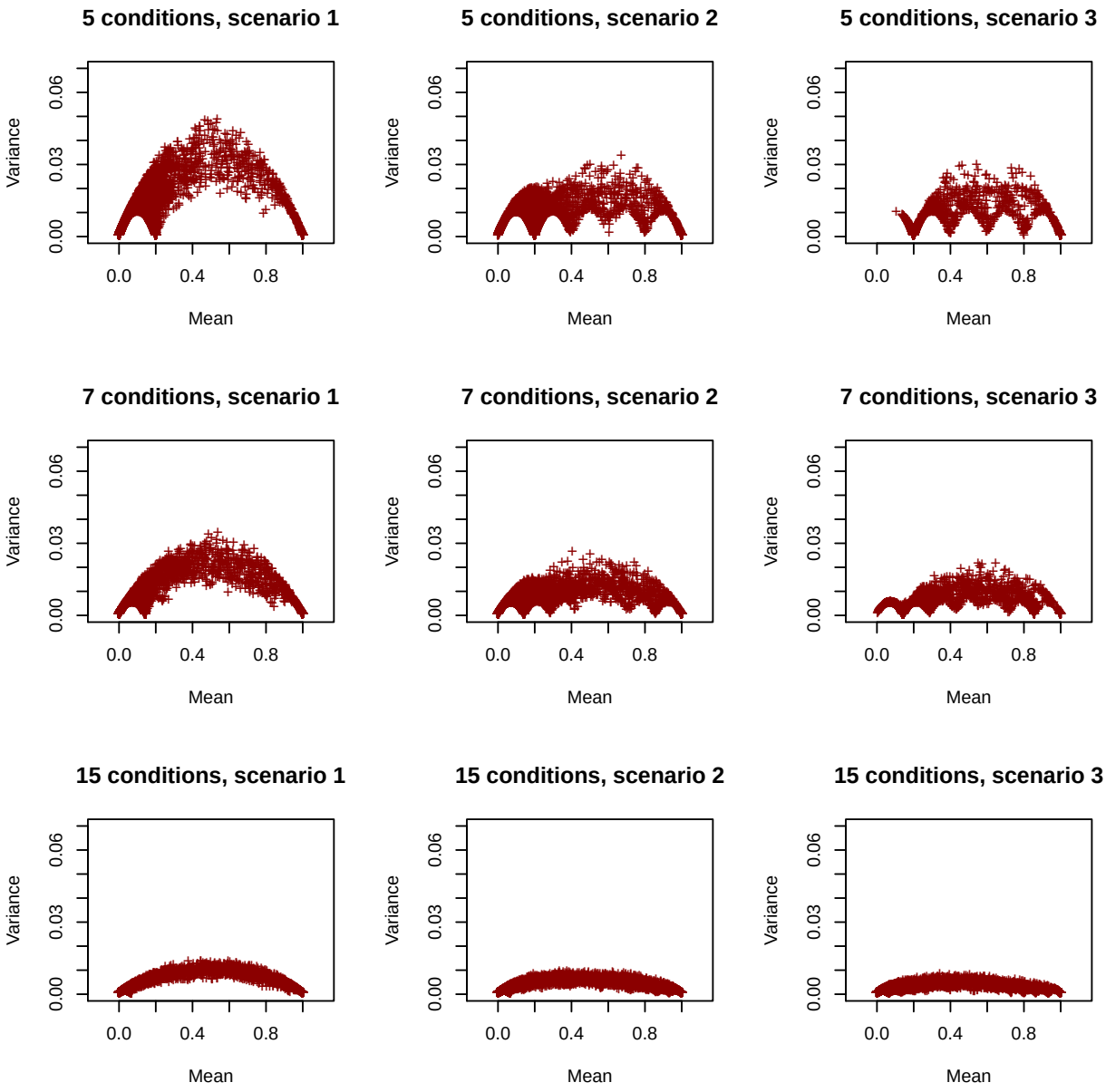

Figure 2: Variance of lambda for nine different scenarios as a function of its mean. The figure illustrates how lambda can achieve precise estimates for intermediate agreements.

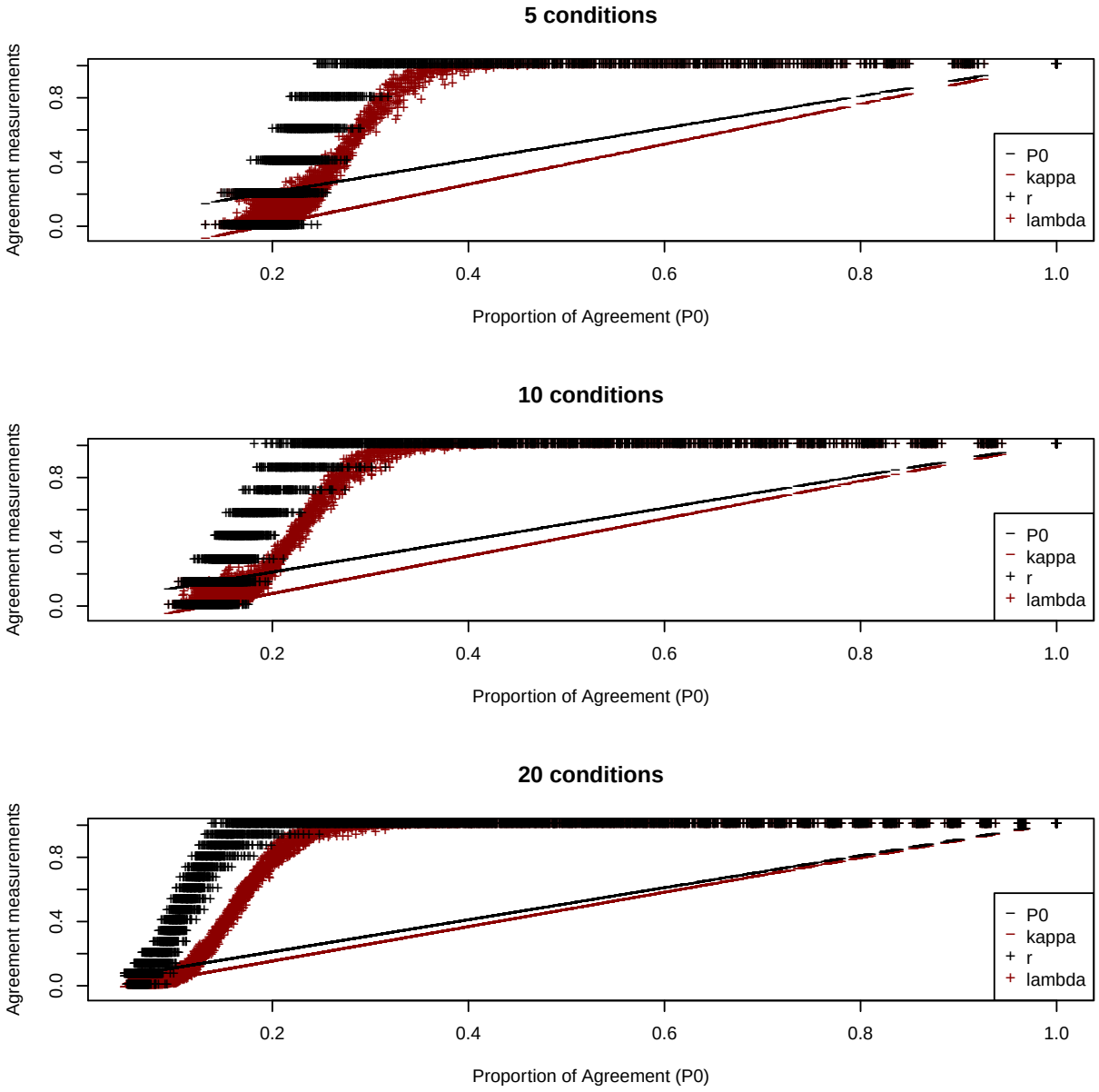

Figure 3: The figure shows the comparison of four agreement measures: P0 (the total fraction of agreement), Cohen's kappa, lambda and r (the total fraction of times the diagonal elements are row and column maxima).

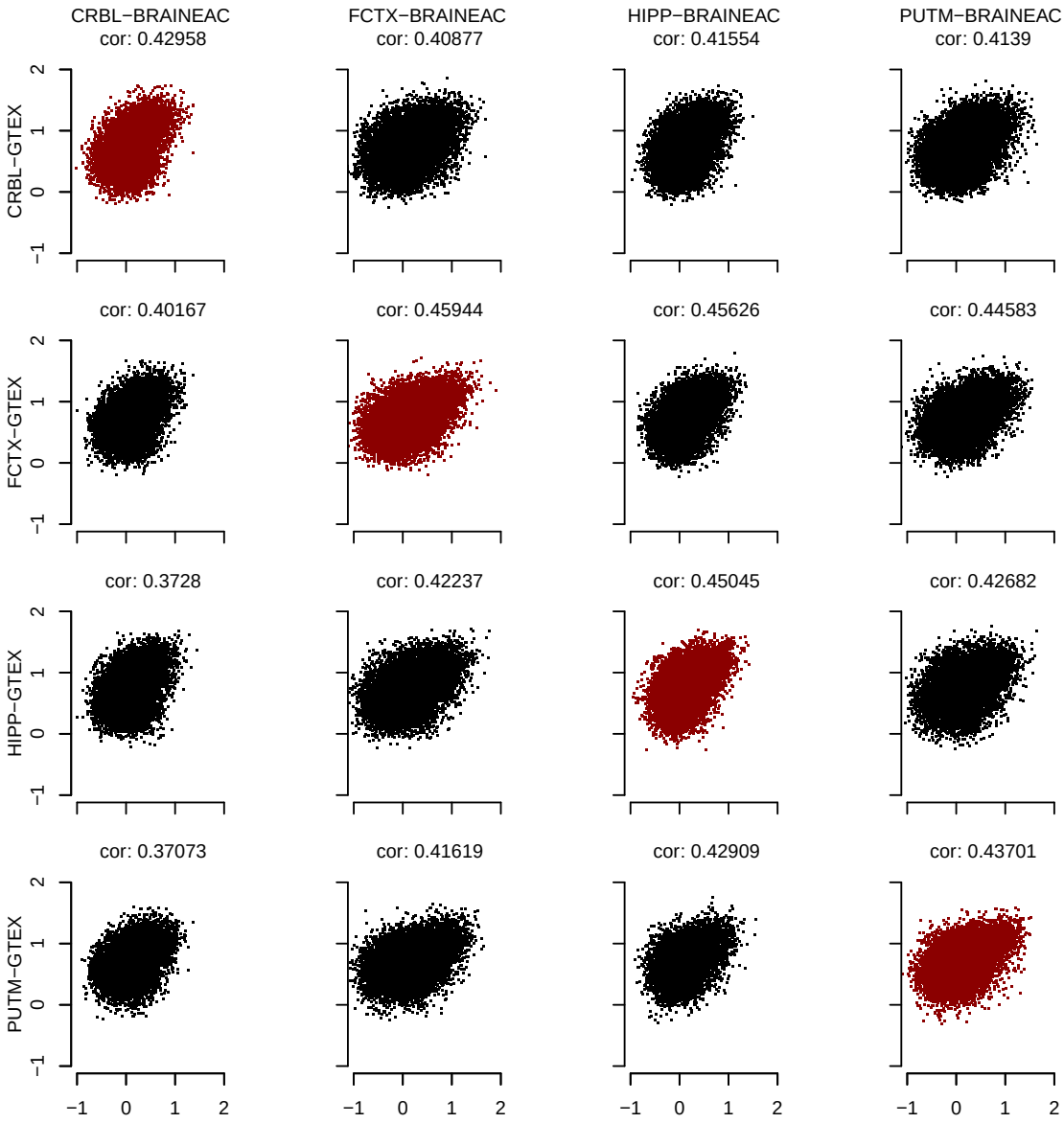

Figure 4: Correlation matrix between networks of four brain tissues across the GTEX and BRAINEAC studies (CRBL:cerebellum, FCTX:frontal cortex, HIPP: hippocampus, PUTM: putamen). The diagonal terms are shown in red. The agreement measure lambda assesses the mean fraction of times the diagonal terms are row and column maxima, given the distribution of the correlation estimates.

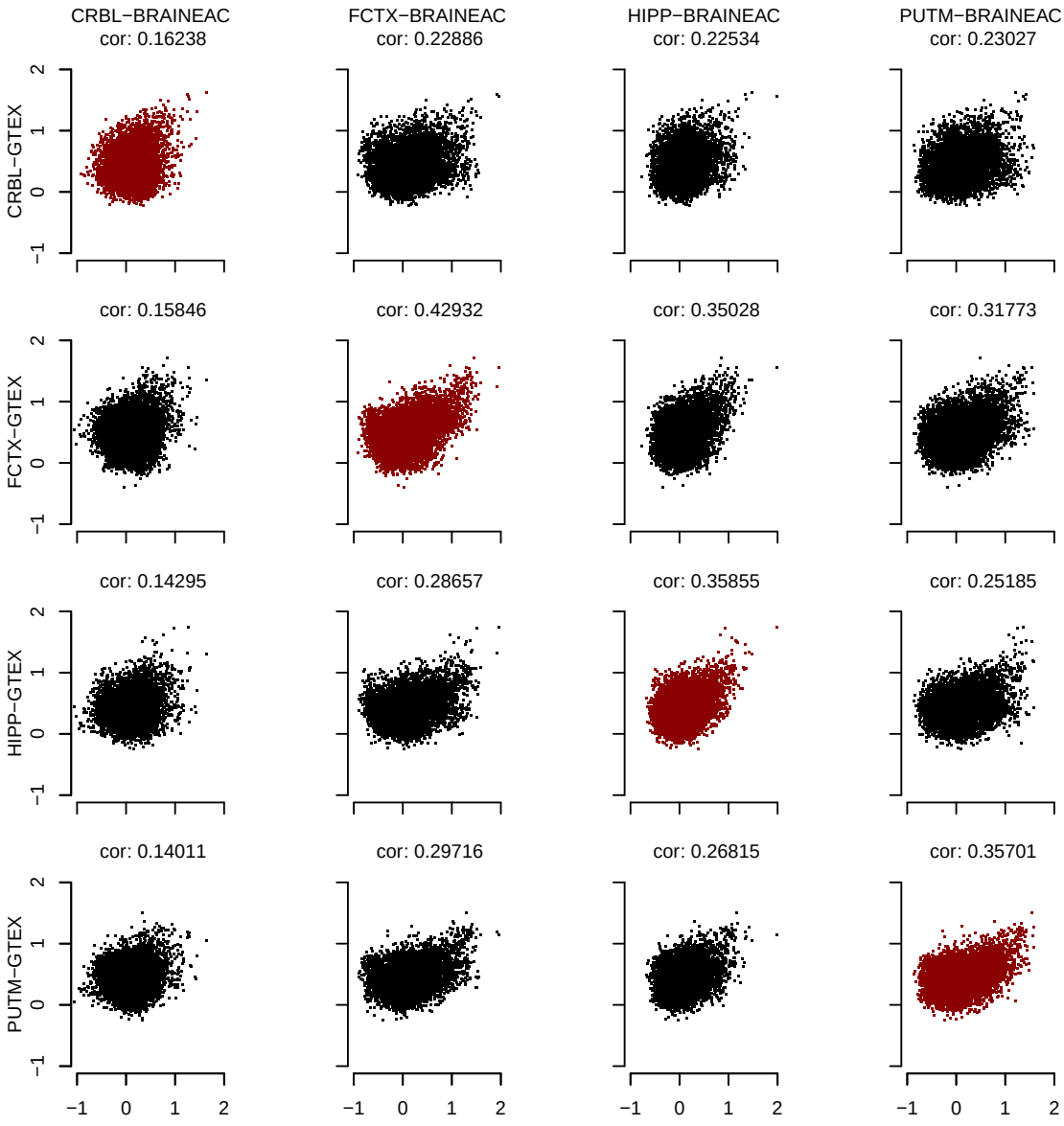

Figure 5: Correlation matrix between networks of four brain tissues across the GTEX and BRAINEAC studies (CRBL:cerebellum, FCTX:frontal cortex, HIPP: hippocampus, PUTM: putamen) for the neuroactive ligand-receptor interaction pathway. The diagonal terms are shown in red.

6 Tables

	CRBL-BRAINEAC	FCTX-BRAINEAC	HIPP-BRAINEAC	PUTM-BRAINEAC
CRBL-GTEX	0.46	0.43	0.44	0.44
FCTX-GTEX	0.43	0.50	0.49	0.48
HIPP-GTEX	0.39	0.45	0.49	0.46
PUTM-GTEX	0.39	0.44	0.46	0.47

Table 5: Z-transformed correlations between GTEX (rows) and BRAINEAC (columns) gene networks in four brain regions. The agreement of the table is lambda=0.5 with vanishing variance due to the large amount of gene pairs involved in the correlations (tens of millions).

	CRBL-BRAINEAC	FCTX-BRAINEAC	HIPP-BRAINEAC	PUTM-BRAINEAC
CRBL-GTEX	0.00015	0.00015	0.00015	0.00015
FCTX-GTEX	0.000148	0.000148	0.000148	0.000148
HIPP-GTEX	0.000148	0.000148	0.000148	0.000148
PUTM-GTEX	0.00015	0.00015	0.00015	0.00015

Table 6: Standard errors of z-transformed correlations between GTEX (rows) and BRAINEAC (columns) gene networks in four brain regions.

	CRBL-BRAINEAC	FCTX-BRAINEAC	HIPP-BRAINEAC	PUTM-BRAINEAC
CRBL-GTEX	4	1	3	2
FCTX-GTEX	1	4	3	2
HIPP-GTEX	1	2	4	3
PUTM-GTEX	1	2	3	4

Table 7: Ranking of network correlations for BRAINEAC (columns) at a given GTEX (rows). The benchmarking of BRAINEAC with respect to GTEX is lambda = 1

lambda	variance	Description	Ref
0.682	0.012	Neuroactive ligand-receptor interaction	hsa04080
0.655	0.024	Nicotine addiction	hsa05033
0.600	0.046	Long-term potentiation	hsa04720
0.579	0.015	Calcium signaling pathway	hsa04020
0.560	0.032	GnRH signaling pathway	hsa04912
0.543	0.038	MicroRNAs in cancer	hsa05206
0.539	0.038	Alcoholism	hsa05034
0.501	0.035	Transcriptional misregulation in cancer	hsa05202

Table 8: Agreement measure lambda between BRAINEAC and GTEX for distinguishability of KEGG pathways in more than 2 tissues out of 4 (lambda > $2/4 = 0.5$)

lambda	variance	Description	Ref
0.858	0.021	Pathways in cancer	hsa05200
0.832	0.014	Calcium signaling pathway	hsa04020
0.805	0.028	Neuroactive ligand-receptor interaction	hsa04080
0.785	0.036	Transcriptional misregulation in cancer	hsa05202
0.757	0.030	Long-term potentiation	hsa04720

Table 9: BRAINEAC benchmarked with respect to GTEX for distinguishability of KEGG pathways in more than 3 tissues out of 4 (lambda > $3/4 = 0.75$)