Documentation brain-coX

Saskia Freytag

26 August 2016

This document outlines the statistical and computational details of all brain-coX's functions. The functions are described in the same order as presented in brain-coX.

brain-coX was built using R and shiny.

1. Loading Data

Currently, brain-coX supports seven large transcriptomic datasets generated from post-mortem human brain samples. Whilst pursuing similar ideas the studies connected to these datasets differed vastly in their design and protocols:

Study	Age Range	Platform	Brains	Samples
Hawrylycz et al	[24 years, 57 years]	Agilent 64K	10	3546
Miller et al	[15 PCW, 21 PCW]	Agilent 64K	4	1134
Kang et al	[4 PCW, 82 years]	Affymetrix Exon 1.0ST	57	1329
Colantuoni et al	[14 PCW, 80 years]	Illumina (custom array)	269	269
Hernandez et al*	[0.5 year, 102 years]	Illumina HumanHT-12	397	908
Trabzuni et al	[16 years, 102 years]	Affymetrix Exon 1.0ST	134	1222
Zhang et al	[22 years, 106 years]	Illumina HumanHT-12 V3.0	101	302

- This dataset contains individuals with abnormal brains. Any number of datasets can be selected for further interrogation.

2. Finding Genes

brain-coX requires three sets of genes to be specified:

- Known disease genes
- Candidate disease genes
- Related disease genes

Known disease genes are genes confirmed to play a role in the disease of interest. Typically, these are all genes identified and validated to be associated with the disease of interest in the literature. Candidate genes are genes that could potentially be involved in the disease pathology but haven't so far been confirmed. These are the genes that the user wants to prioritise. In a next generation sequencing study, these are all genes found to habour de novo mutations in individuals affected with the disease of interest. Related disease genes are genes that are known to be associated with related diseases. For example, if the disease of interest is epileptic encephalopathy then related disease genes would include all genes known to play a role in general epilepsy. While it is not necessary to define related disease genes, we strongly recommend their use. Note that related disease genes only inform the cleaning step by preventing interesting signal from being removed, but they do not inform the prioritisation step.

brain-coX assumes HUGO gene identifiers. When there are multiple aliases for the same gene, brain-coX automatically converts these according to the USCS genome browser. The user is informed about gene name conversions as well as genes that cannot be found in the datasets in the output.

3. Cleaning

brain-coX applies either removal of unwanted variation (RUV, Gagnon-Bartsch and Speed) or conventional data cleaning (background correction and quantile normalization) to all datasets separately. The RUV cleaning procedure is data-driven and adaptively removes systematic noise while also taking the user's research interest into account. This allows the combination of datasets as well as accurate estimation of gene-gene correlations (see Freytag et al).

The main principle behind RUV is the use of negative control genes in order to estimate the effect of systematic noise on the gene expression measurements of each gene. Negative control genes are genes that are affected by systematic noise, but, crucially not by any biological variation of interest. The user can choose between the use of recommended housekeeping genes (adapated from Eisenberg and Levanon) and empirically chosen genes as negative control genes. brain-coX automatically excludes all known, candidate and related disease genes from being negative control genes. Note that only the number of empirically chosen negative control genes can be specified, however brain-coX will inform the user about the number of housekeeping genes used in each dataset in the output. Moreover, brain-coX uses slightly different values for the regularization parameters depending on which set of negative control genes are used. These can be found in the following table

Study	k	nu - Housekeeping Genes	nu - Empirically Genes
Hawrylycz et al	5	25000	50000
Miller et al	4	500000	500000
Kang et al	3	15000	10000
Colantuoni et al	3	35000	25000
Hernandez et al	1	0	20000
Trabzuni et al	4	250000	350000
Zhang et al	1	750	500

4. Prioritisation

Gene priortisation is the core function of brain-coX and we have implmented a proven and successful prioritisation strategy, BrainGEP. We have extended BrainGEP to be period-specific, i.e. the user can subset datasets in order to include only samples from periods of interest.

BrainGEP:

The BrainGEP prioritising strategy is described in detail by Oliver et al. brain-coX implements this strategy for every selected dataset individually with some notable improvements. In particular, we developed a new approach to set the threshold that defines at which absolute correlation value interactions are declared significant. Briefly, we set the threshold for an absolute value of the Pearson correlation coefficient that corresponds to the top user selected proportion of ranked random genes. These random genes are ranked according to their maximum absolute correlation with any of the known disease genes. This process is repeated 1000 times and the average maximum absolute correlation was calculated and set as a threshold. Note that the size of the randomly selected genes corresponds to the number of candidates in the various datasets. Unlike, BrainGEP the Pearson correlation coefficients are calculated in a weighted manner, where samples are weighted by the inverse of the square root of the number of samples in their corresponding brain.

brain-coX also tries to interpret the significance of the prioritisation results. In particular, brain-coX calculates the expected overlap of proritised genes between all selected datasets, i.e. how many genes should be expected to overlap by chance given the number of prioritised genes in each selected dataset and the total number of candidate genes. Note that this result is only useful if assuming that genes priortised in several datasets are more likely to be genuinely involved in the disease pathogenesis. This is necessary to ensure that each brain

is contributing roughly in an equal manner to the calculated correlation despite the very differing numbers of samples in each brain.

brain-coX also offers an alternative prioritization approach:

geneRecommender:

Unlike in the BrainGEP prioritisiation strategy, brain-coX applies geneRecommender to the dataset combining all selected datasets. Datasets are combined after they are individually cleaned and rescaled to prevent different scaling from driving correlation results. The geneRecommender algorithm is described in detail by Owen et al. Briefly, the algorithm first identifies a subset of all samples that is most informative with regards to the behaviour of the known disease genes. It then uses these samples to score the candidate genes according to how similar their behaviour is to the behaviour of the known genes. The calculated score can then be ranked resulting in a ranking of the candidate genes from most interesting to least interesting. The user can select the number of genes that should be displayed. The user is informed about how many samples are deemed informative by geneRecommender in the output. It is possible that geneRecommender only uses a very small subset of all samples. In this situtation it might be advisable to force geneRecommender to use all samples, which can be done using the checkbox "*Use all samples*".

5. Visualisation

brain-coX also provides visualisation for gene-gene interactions between known, candidate and priortised disease genes as well as user specified genes as obtained using datasets individually or in a combined fashion.

Networks:

The visualised networks are based on the BrainGEP approach as well as periods specified therein. However, the proportion of ranked random genes can be independently selected as well as the datasets to be included in the visualisation. For each interaction, brain-coX visualises in how many selected datasets this particular interaction is observed by the weight of edges in the network graph. Furthermore, brain-coX also visualises whether all datasets agree on an interaction being an inhibition or activation or whether there is disagreement between the selected datasets using the colour of the edges.

Note that the layout of the network can be changed to resemble a tree structure.

Correlations:

This visualsation option produces correlation plots based on weighted Pearson correlation coefficients calculated for a selected developmental period. The correlation is calculated only using samples that are from individuals in the selected developmental period. Samples are weighted by the inverse of the square root of the number of samples in their corresponding brain. This is necessary to ensure that each brain is contributing roughly in an equal manner to the calculated correlation despite the very differing numbers of samples in each brain. Note that partial correlations can also be calculated but these are not weighted.

There are several plot output options for the correlation plot. The genes in the correlation plot can be ordered according to how they cluster, the first principle component or the original order of the genes can be maintained. Finally special genes of interest can be highlighted on the axes.

6. Analysis

The user selects both the genes and the sets of periods of interest. Using these brain-coX will render a correlation plot, where the lower triangle displays the correlations of the genes as estimated from samples of

individuals in the first selected set of periods and the upper triangle displays the correlations as estimated from samples of individuals in the second selected set of periods. brain-coX also tests whether individual correlations (using Fisher's test for correlations appropriately adjusted for the number of tests performed), as well as the set of correlations, are significantly different between the two sets of periods (using a covariance test). The correlations are calculated by combining datasets and rescaling them as well as weighting proportionally to the number of samples in each brain.

brain-coX helps the user to interpret the output by providing some interactive features on the plot. By clicking on the gene lables of the y-axis all other gene-gene correlations are partially masked except for the ones relevant to the selected gene. Clicking anywhere but on the y-axis will re-establish the original plot.

7. Hot Candidate

brain-coX allows the user to really focus on a particularly interesting candidate gene. There are two different investigation avenues.

Networks:

The user selects the developmental period of interest. Weighted correlations are then calculated using all samples that fall within the specified period. Thus, this approach combines datasets (rescaling is performed). Interactions between genes only result in an edge on the network graph when their absolute value is above the user specified threshold.

Analysis:

brain-coX is also able to plot the change of gene-gene correlations specific to the selected hot candidate over time, i.e. in different periods. Note that periods with less than 10 samples are not plotted.

8. Investigation

This option allows you to investigate the power of your known genes for prioritising by comparing the empirical cumulative distribution function (ECDF) of the absolute correlations between the known genes to the ECDF of the absolute correlations between a set of random genes. If the ECDF of the kown genes demonstrates that the absolute correlations between known genes are generally higher than the absolute correlations of a set of random genes (the ECDF of the known genes lies below the ECDF of the random genes) then power for prioritising candidates is expected to be high. Correlations are not calculated by using an unweighted approach. Note that this is only an indication of the expected power; actual results might differ substantially.

Periods

The developmental periods used in brain-coX were defined by Kang et al and consist of 15 different periods ranging from embryonic to late adulthood. Note that Y denotes years and PCW is an abbreviation for post conception weeks.

Period	Description	Age Range
1	Embryonic	4-8 PCW
2	Early fetal	8-10 PCW
3	Early fetal	10-13 PCW
4	Early mid-fetal	13-16 PCW

Period	Description	Age Range
5	Early mid-fetal	16-19 PCW
6	Late mid-fetal	19-24 PCW
7	Late fetal	24-38 PCW
8	Neonatal and early infancy	Birth-6 M
9	Late infancy	6 M-1 Y
10	Early childhood	1 Y-6 Y
11	Middle and late childhood	6 Y-12 Y
12	Adolescence	12 Y-20 Y
13	Young adulthood	20 Y-40 Y
14	Middle adulthood	40 Y-60 Y
15	Late adulthood	60 Y +

Trouble Shooting

Why am I obtaining an error message saying "*Something isn't right here!*" ?

> At the moment brain-coX is unable to unload datasets. So the best way to get around your problem is to restart the application and load only the datasets that you want to use for the prioritisation.

I am waiting a long time for my analysis to be finished. How can I tell whether brain-coX is still working?

> brain-coX deals with very large datasets, so please be patient. You can see the progress brain-coX is making in the upper right hand corner. If you are frequently using our tool you might want to obtain a version of brain-coX that runs locally. In that case please contact Saskia Freytag via freytag.s@wehi.edu.au.

None of the answers above addresses my problem. Where can I get help?

> Please address any questions to Saskia Freytag via freytag.s@wehi.edu.au. Of course any feedback and problem reporting are also welcome.