Consensus modules: modules present across multiple data sets

Peter Langfelder and Steve Horvath

Eigengene networks for studying the relationships between co-expression modules.

BMC Systems Biology 2007, 1:54

Why consensus modules?

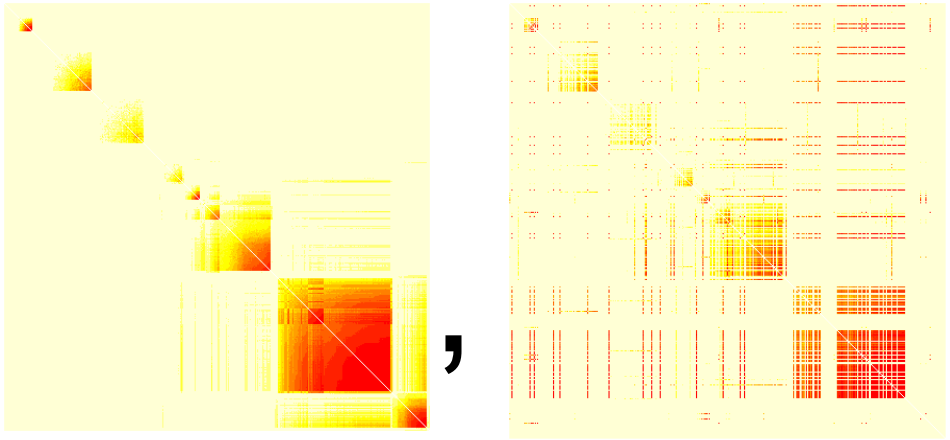
- Given several independent data sets, find modules that are present in all (or a specified majority) of the data sets
- Rationale
  - Find co-expression patterns common to multiple studied conditions
  - Find common, robustly defined modules across several independent data sets (e.g., from GEO) that study the same conditions

Finding consensus modules

- Modules group together densely interconnected genes
- Consensus modules group together genes densely connected in all conditions
- Our solution: find the consensus gene-gene similarity and use it with clustering to find modules

Finding consensus modules

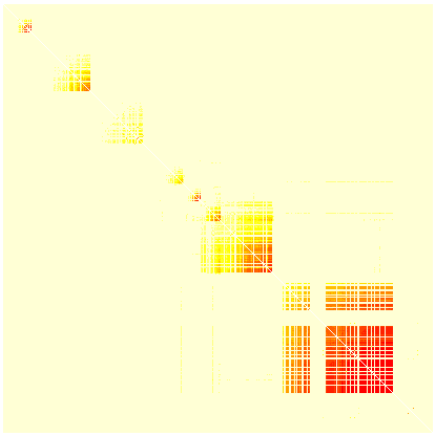
- Calibrate input networks to make them comparable



Network 1 Network 2

Finding consensus modules

- Calibrate input networks to make them comparable
- For 2-3 sets: Take component-wise minimum



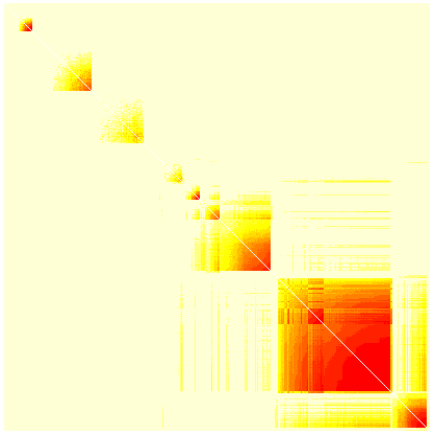
Consensus

Component-wise

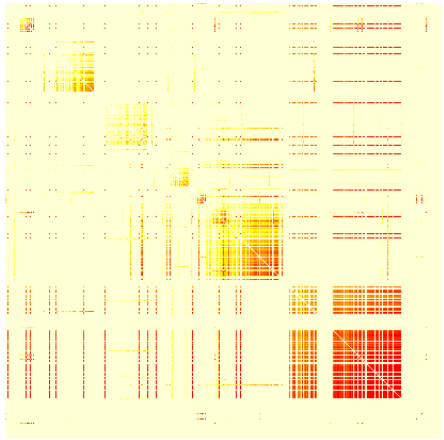
= min

pmin()

(



,



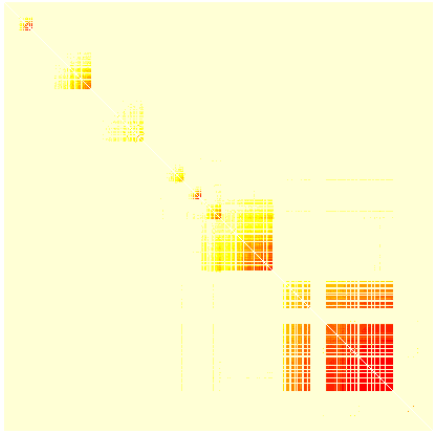
)

Network 1

Network 2

Finding consensus modules

- Calibrate input networks to make them comparable
- For 4 sets or more: suitable quantile (for example, quartile)

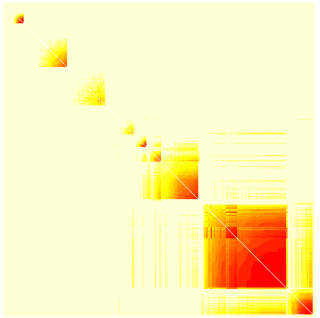
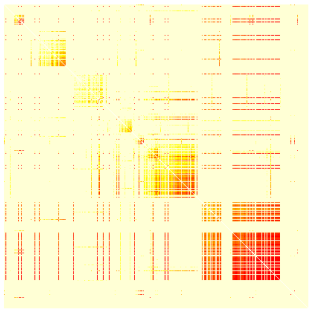
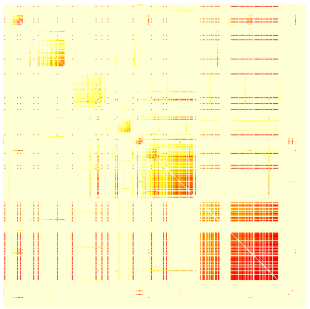
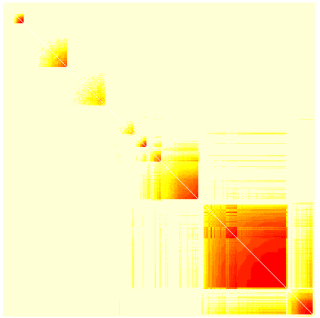


Consensus

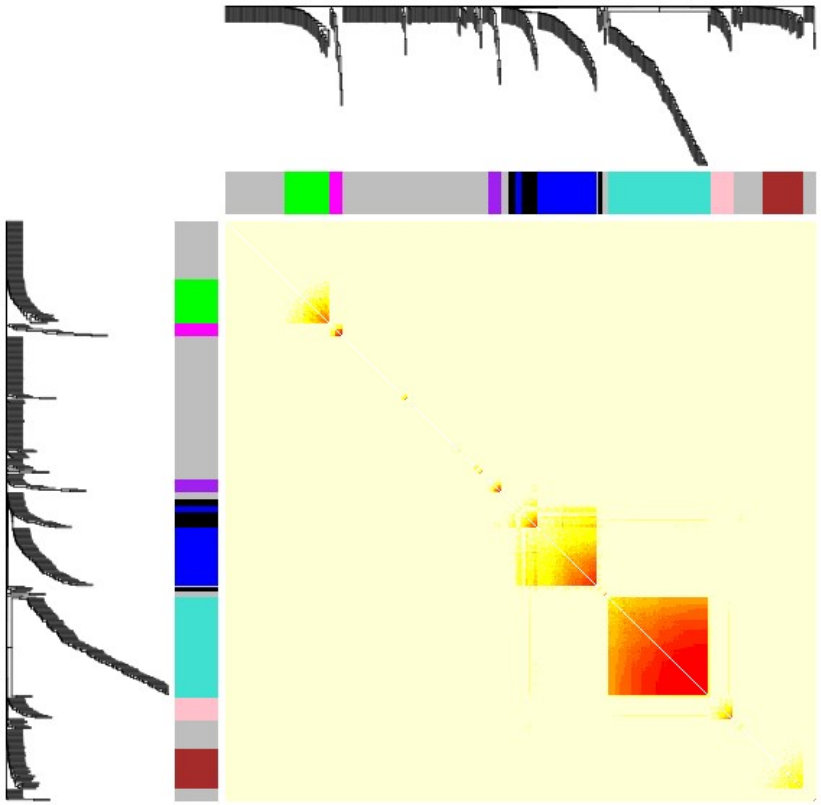
=

component-wise quantile

`pquantile()`

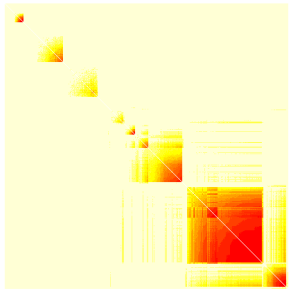
( Network 1 ,  Network 2 ,  Network 3 ,  Network 4 ,...)

Finding consensus modules

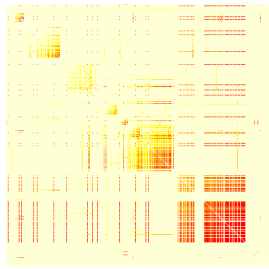


Consensus

=min(



,



)

Network 1

Network 2

Consensus modules are defined from clustering of consensus similarity

R implementation: `blockwiseConsensusModules`

- Input:
  - Expression data in "multi-set" format
  - Options for splitting data into smaller blocks if there are too many genes to be handled in one block ("blockwise")
  - Network construction options for constructing individual networks
  - Network calibration options
  - Consensus quantile

R implementation: `blockwiseConsensusModules`

- Output:
  - Consensus module labels,
  - Gene clustering tree (or trees if the data was split into blocks)
  - Module eigengenes
  - Other diagnostic output
- Introductory tutorial: Consensus analysis of female and male liver expression data (Tutorial II) at

labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials

Consensus modules vs. Module preservation statistics

- Consensus modules are by construction present (i.e., preserved) in all (or most) input data sets
- If a module identified in a reference data set is strongly preserved in test data set(s), it would also be a consensus module among the reference and test sets
- Consensus module construction treats all data sets the same; module preservation statistics require a reference and a test data set(s)
- Consensus modules are best suited to answer a different set of questions than module preservation statistics

Application 1: Consensus modules across multiple lung cancer data sets

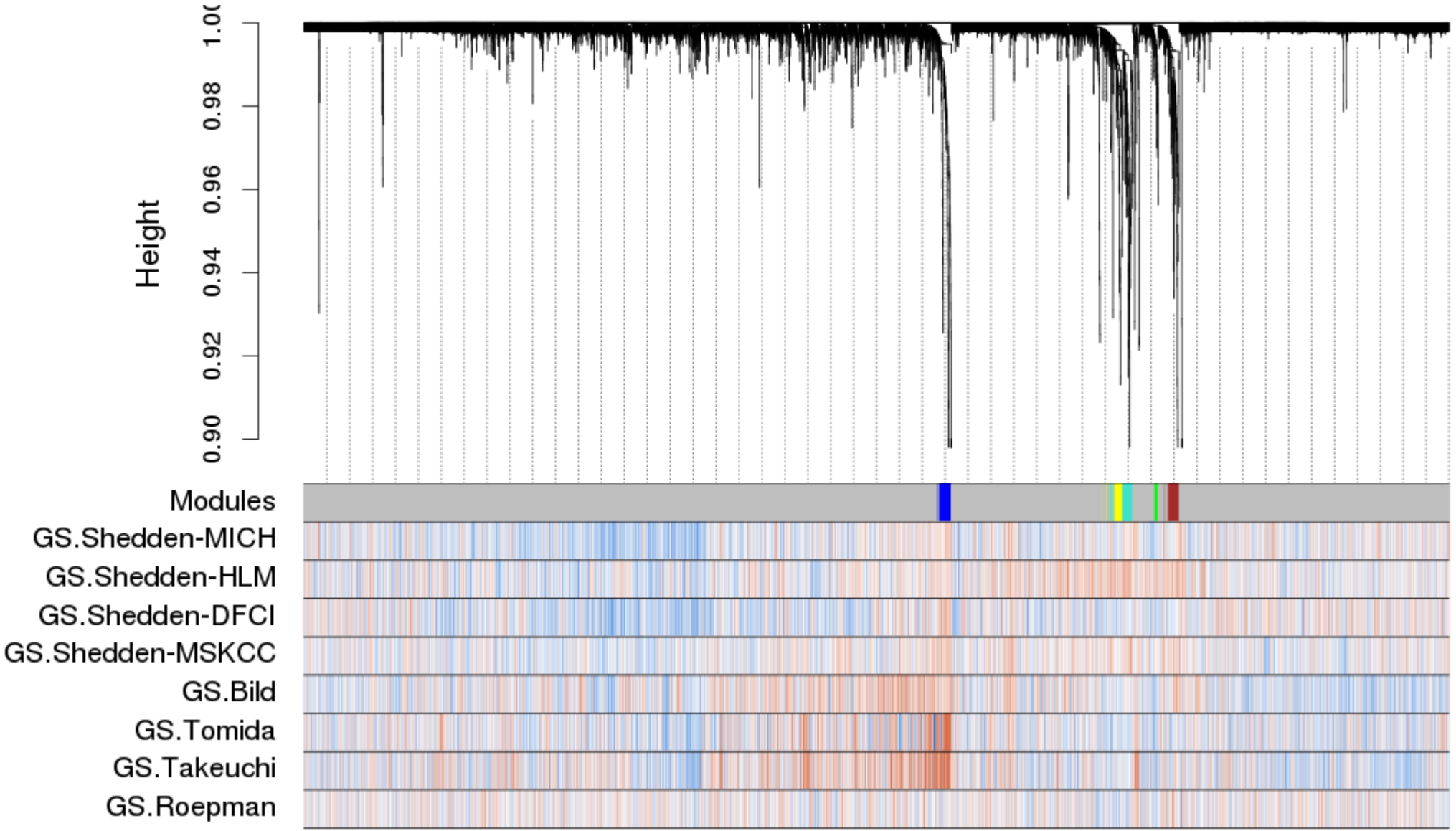
Eight publicly available lung cancer sets

- 4 independent data sets described in Shedden et al, *Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nat Med. 2008 Aug;14(8):822-7. Epub 2008 Jul 20. (Affy U133A)
- Bild et al, *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.* Nature. 2006 Jan 19;439(7074):353-7 (Affy U133plus2)
- Tomida el al, *Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis.* J Clin Oncol 2009 Jun 10;27(17):2793-9. (Agilent-014850)
- Takeuchi et al, *Expression profile-defined classification of lung adeno-carcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors.* J Clin Oncol. 2006 Apr 10;24(11):1679-88 (Agilent 21.6K custom array)
- Roepman et al, *An immune response enriched 72-gene prognostic profile for early-stage non-small-cell lung cancer.* Clin Cancer Res. 2009 Jan 1;15(1):284-90 (Agilent-012391)

Eight publicly available lung cancer sets

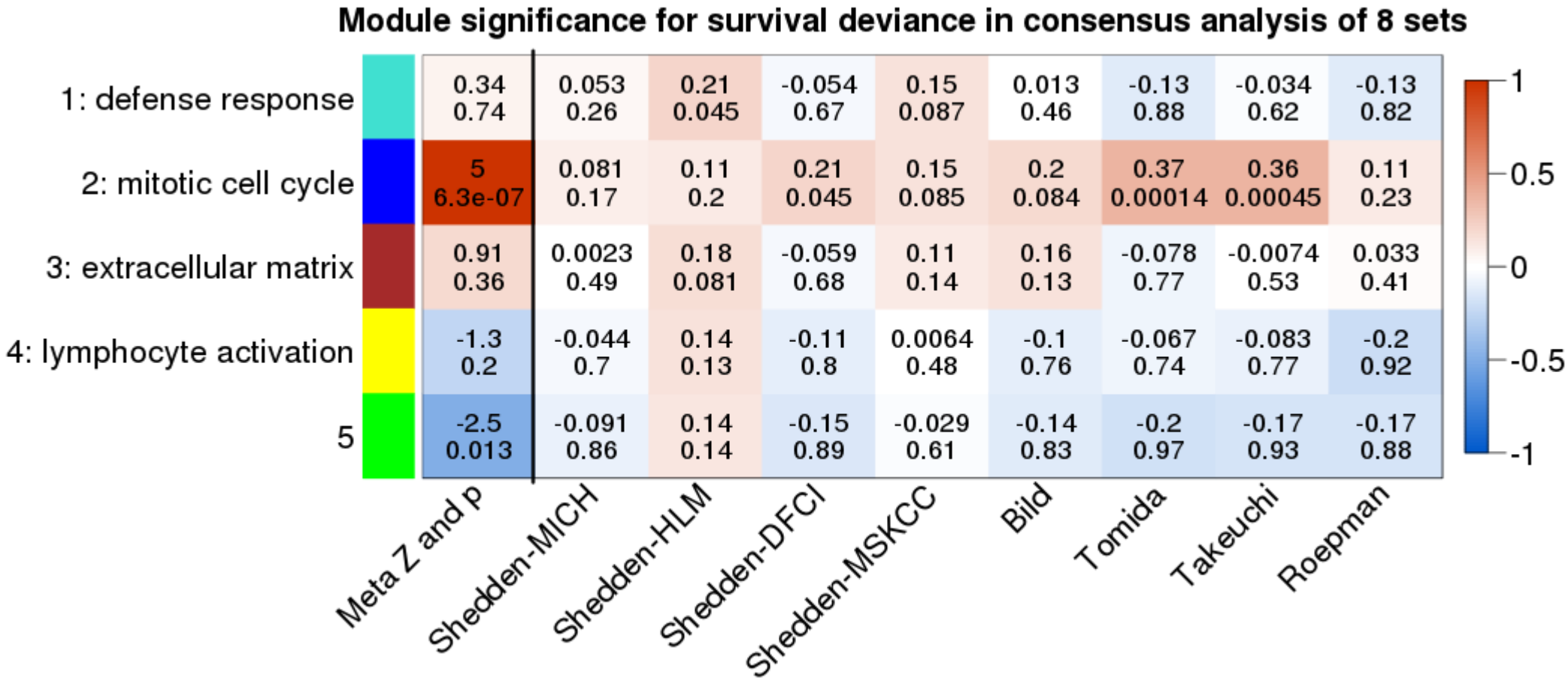
- Concordance between the sets is poor
- Difficult to find genes consistently related to survival time
- For consistency: restrict analysis to adenocarcinoma
- Since we have 8 sets, we use the quartile instead of the minimum in consensus network construction

Consensus modules across 8 data sets



Red: Upregulated in patients with short survival time
Blue: Downregulated in patients with short survival time

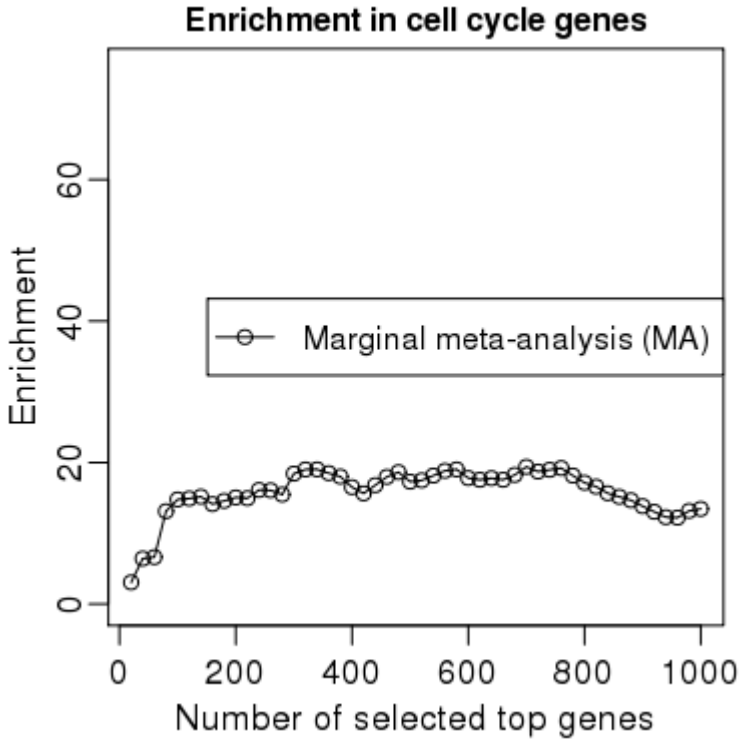
Association of consensus modules with survival deviance



- Cell cycle module shows weak but consistent association with survival time (meta-analysis p-value = 6e-7, highly significant)
- Reflects the known fact that fast-proliferating cancers indicate a poor prognosis for patient

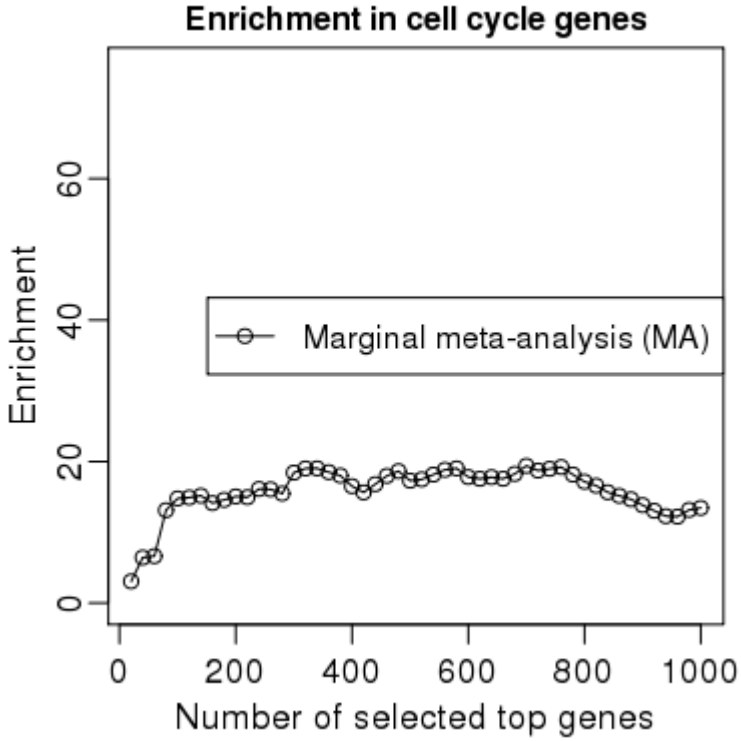
Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment



Using consensus modules to study genes related to a clinical trait

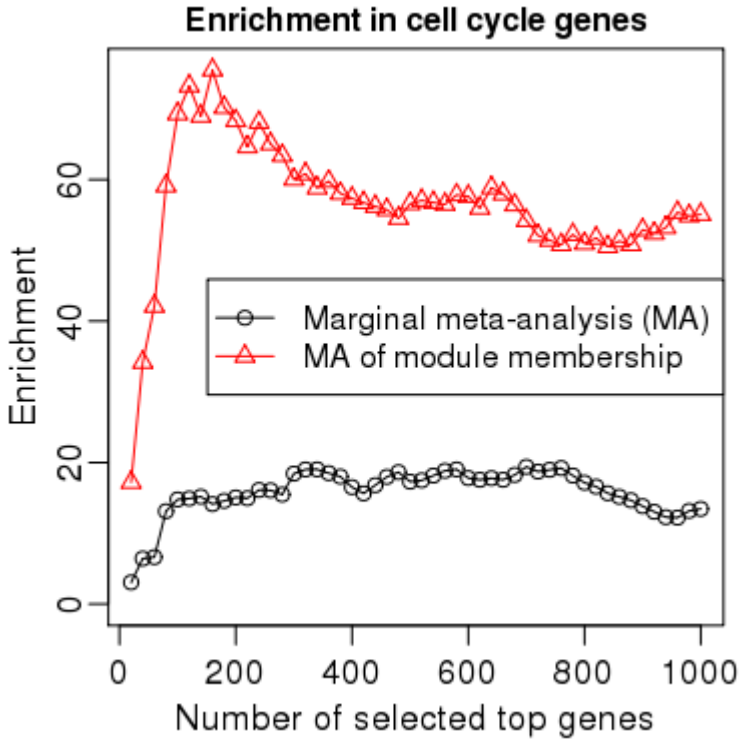
- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment



- Can also study enrichment of genes with highest connectivity in survival-associated module

Using consensus modules to study genes related to a clinical trait

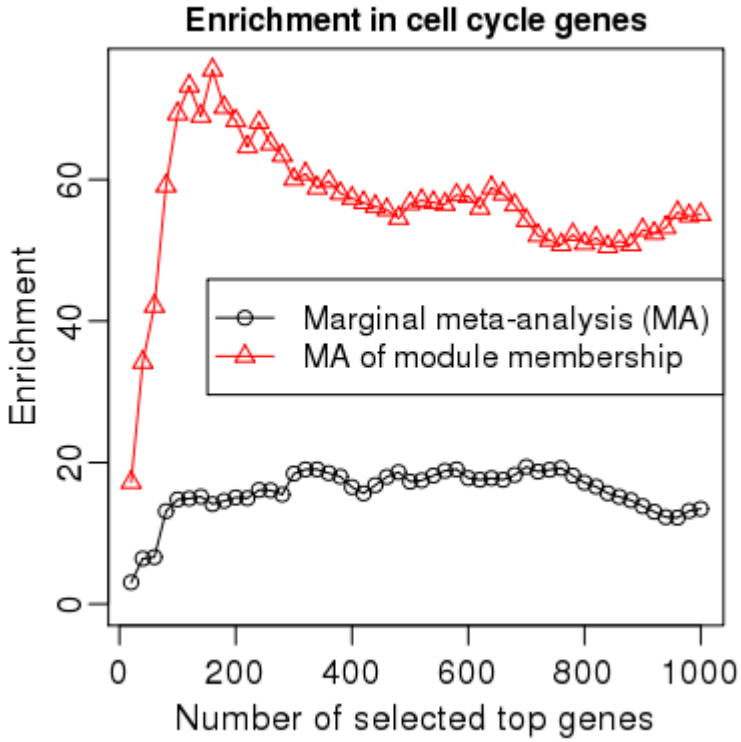
- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment



- Can also study enrichment of genes with highest connectivity in survival-associated module
- Find much higher enrichment
- Consensus module analysis can lead to better biological insights when pooling several gene expression data sets

Using consensus modules to study genes related to a clinical trait

- Challenge: how to identify functional categories that associate with a trait (survival time)
- Can use meta-analysis to select genes related to survival time across all data sets, then study their enrichment



- Can also study enrichment of genes with highest connectivity in survival-associated module
- Find much higher enrichment
- Consensus module analysis can lead to better biological insights when pooling several gene expression data sets

Application 2: Consensus modules across 4 brain regions in Huntington's Disease patients and controls

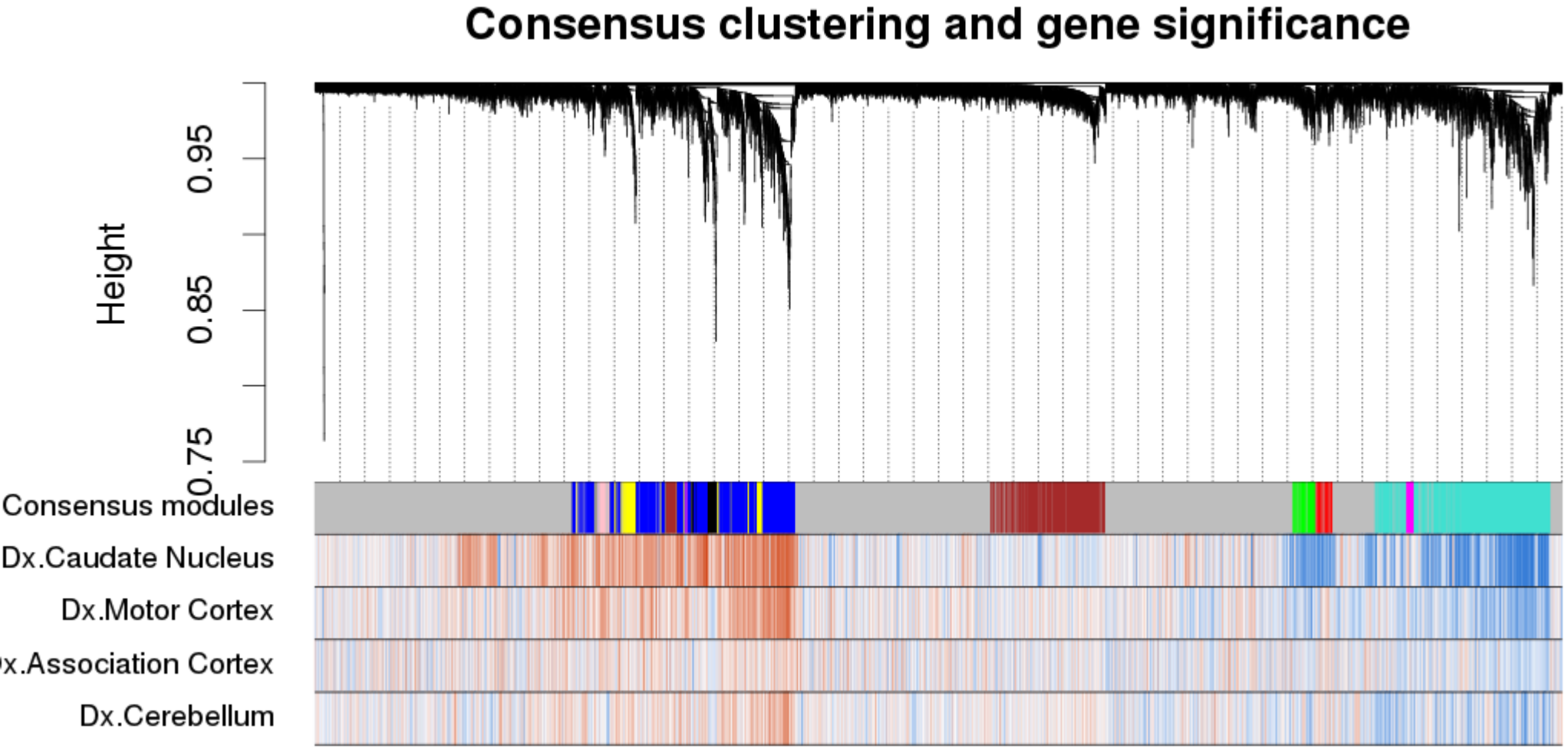
Data

- Huntington's disease primarily affects motor skills
- Biggest changes are observed in Caudate Nucleus (CN), much smaller changes in Cortex and Cerebellum (CB)
- Disease causes dying of neurons and increase in astrocytes/ oligondendrocytes (inflammatory response)
- Hodges et al (2006): Measured expression in Caudate Nucleus, Motor Cortex, Association Cortex, Cerebellum in HD patients and controls
- Here: consensus analysis of the 4 data sets

Hodges A et al, *Regional and cellular gene expression changes in human Huntington's disease brain.* Hum Mol Genet. 2006 Mar 15;15(6):965-77

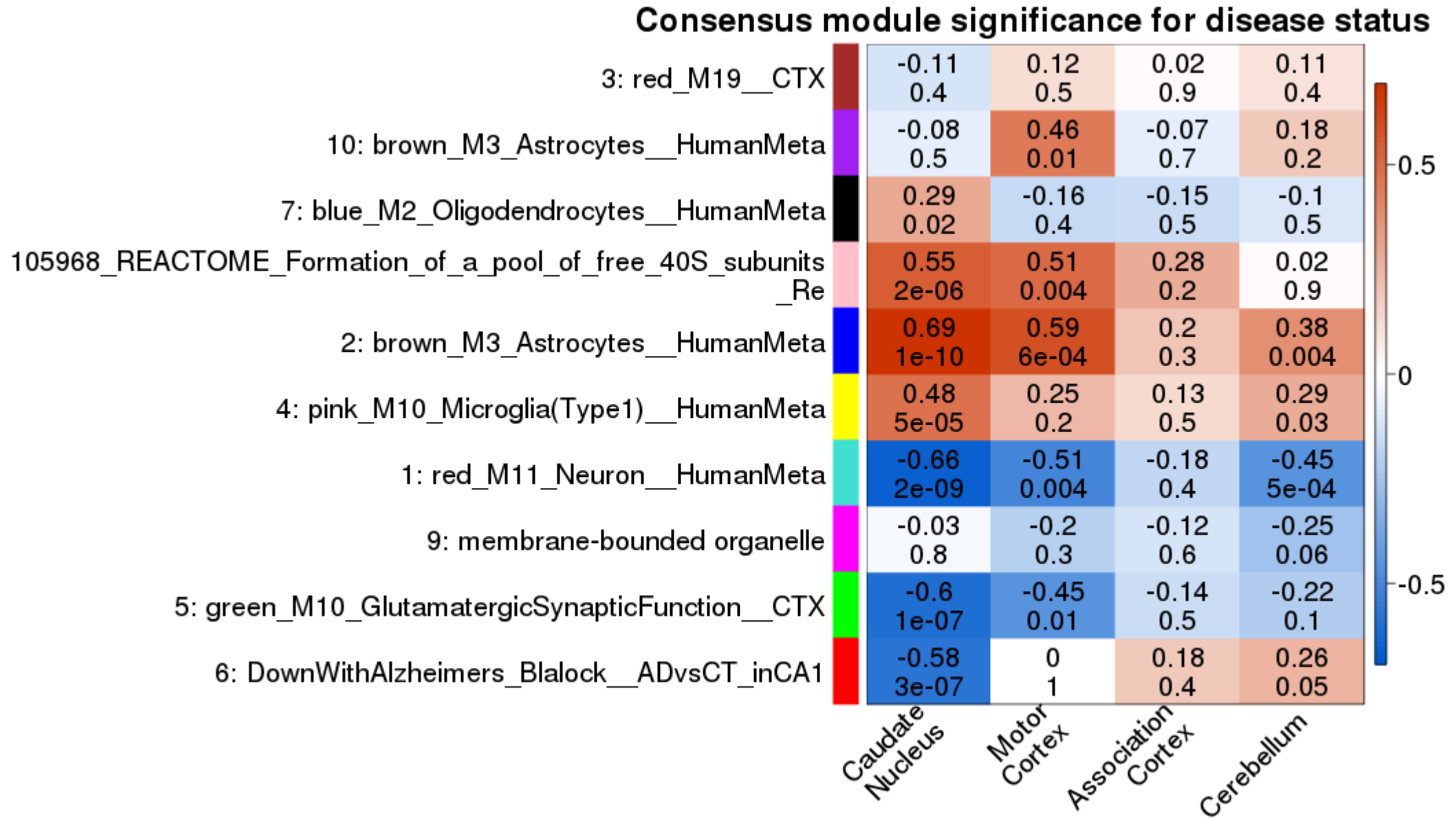
Consensus network and modules

- Two large branches that correspond to neurons and astro/oligodendrocytes



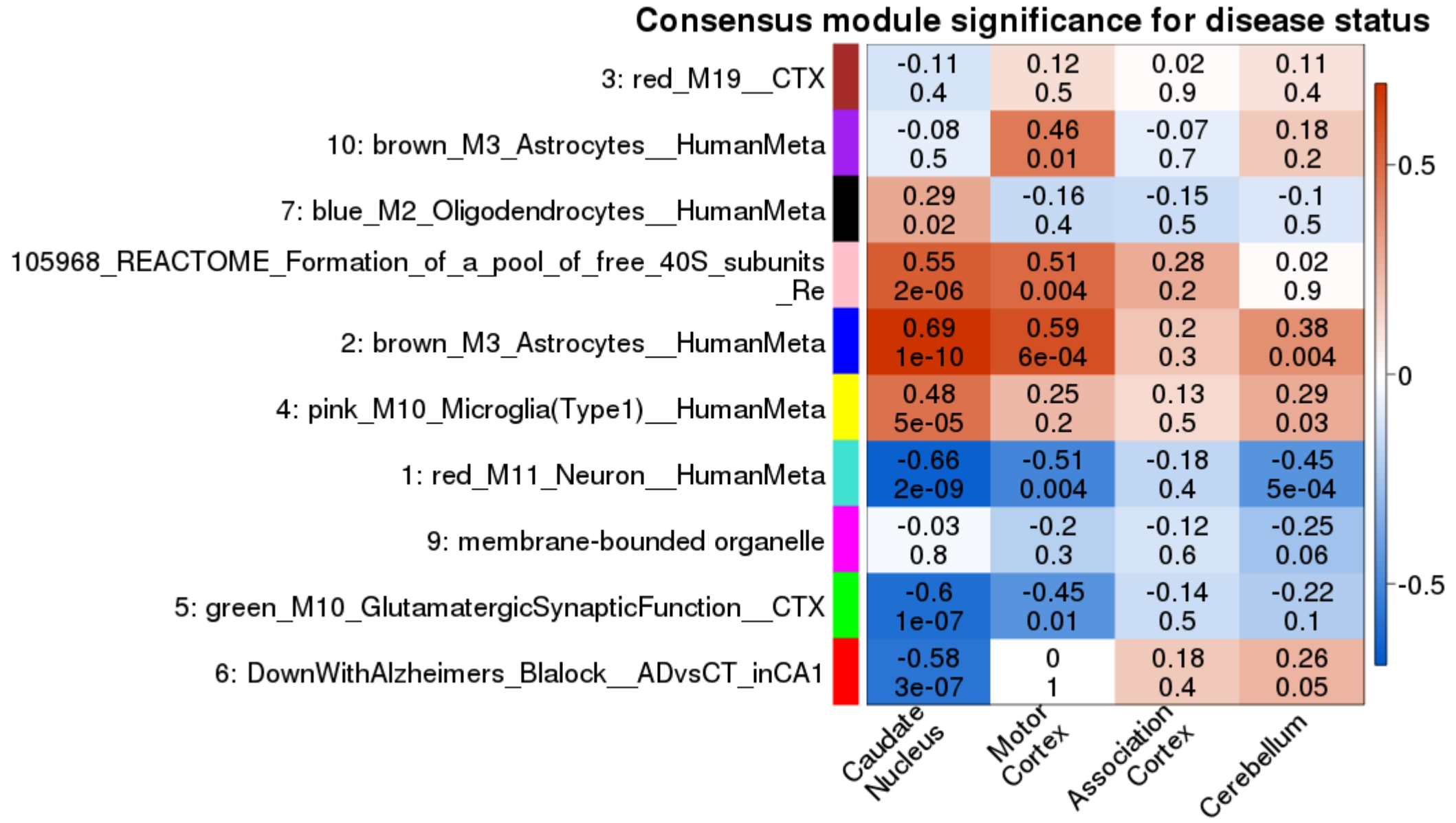
- Red: underexpressed in HD; blue: overexpressed in HD

Significance of modules for disease status



- Message: The disease effect is strongest in Caudate Nucleus and Motor Cortex, consistent with HD manifestations

Significance of modules for disease status



- The modules relate to disease status similarly across all tissues

Study meta-networks built from eigengenes of consensus modules

- Recall: modules are represented by their eigengenes (singular vectors obtained from SVD)
- Each consensus module has one eigengene in each data set
- In each data set: correlation among module eigengenes gives a bird-eye view of the entire gene network: a correlation matrix of 12k genes is reduced to a correlation matrix of 10 eigengenes
- Correlation of eigengenes reflects how the underlying pathways, processes, cell types, etc work together
- It may be interesting to study how eigengene correlation changes between data sets

Langfelder and Horvath, *Eigengene networks for studying the relationships between co-expression modules*. BMC Systems Biology 2007, 1:54

Preservation of eigengene networks between brain regions

- Eigengene network defined as a signed network with power β=1:

$$A_{ij} = \frac{1 + cor\left(E_i, E_j\right)}{2}$$

- Preservation network: measures how much eigengene correlation varies among data sets

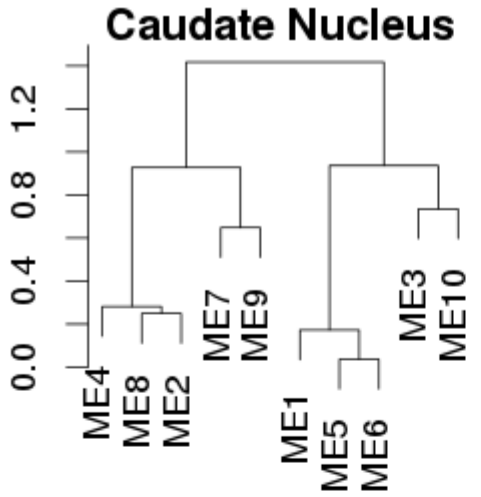
$$Pres_{ij}^{(1,2,\ldots)} = 1 - \left[max\left(A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots\right) - min\left(A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots\right) \right]$$

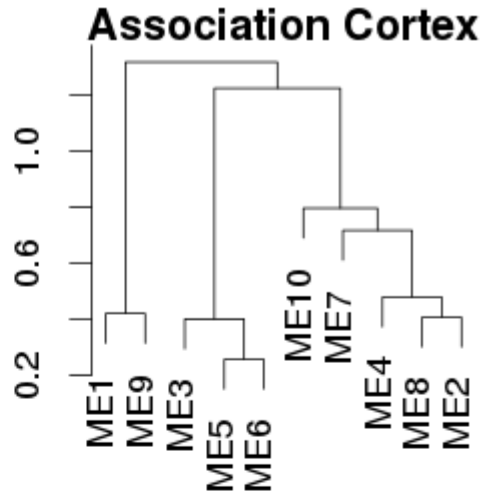
- Mean preservation: measures overall preservation of eigengene networks among data sets

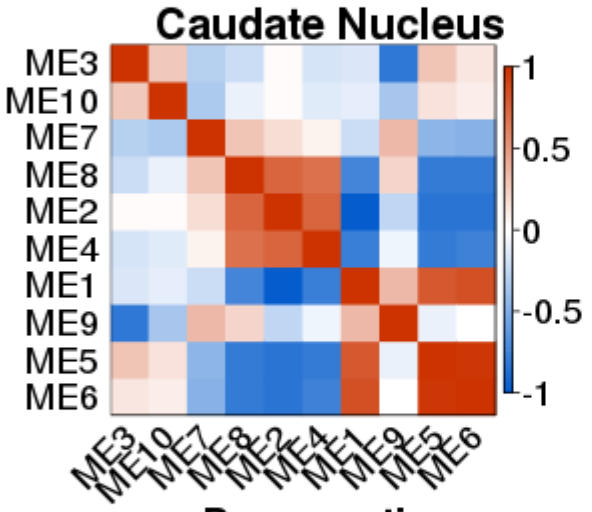
$$D^{(1,2,\ldots)} = mean_{i<j} P_{ij}^{(1,2,\ldots)}$$

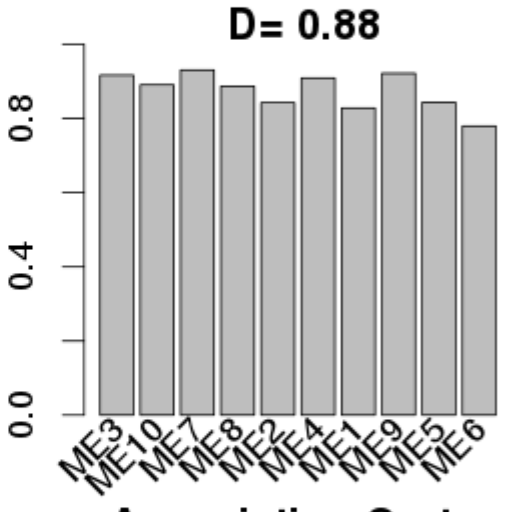
Preservation of eigengene networks between CN and Association CTX

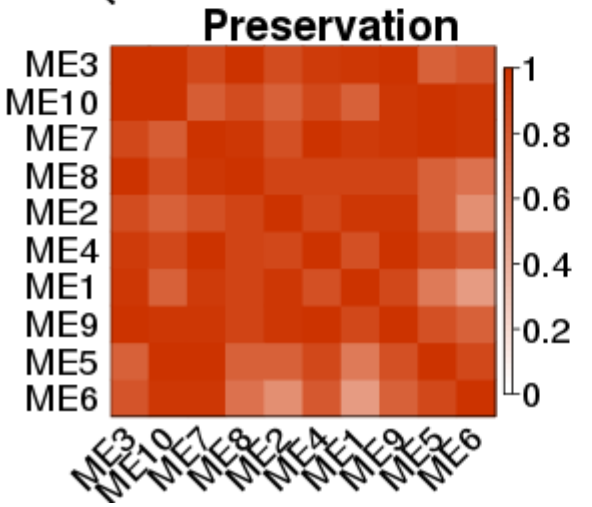
- Preservation is relatively high: D=0.88
- Preservation of networks among neuronal modules is lower than the networks among inflammation-related modules
- We did not discover a cure for HD, but certainly have "Food for thought" for biologists

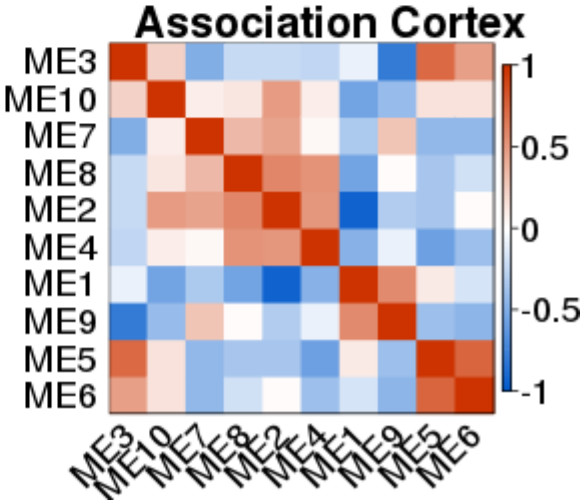












What have we learned?

- Cancer application: consensus analysis identifies a module consistently related to survival time
- The module provides a cleaner biological interpretation than genes identified using standard meta-analysis
- Huntington's disease application: Consensus module analysis shows that HD affects different brain regions in a broadly similar manner but also shows differences in the way the regions are affected
- Consensus eigengene network analysis provides a way to study commonalities and differences in network organization of gene expression or other genomic data

References

- Consensus modules and eigengene networks:

Langfelder P, Horvath S (2007)
Eigengene networks for studying the relationships between co-expression modules.
BMC Systems Biology 1:54
labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/EigengeneNetwork/

- Cancer application of consensus modules:

Langfelder P, Mischel PS, Horvath S (2013)
When Is Hub Gene Selection Better than Standard Meta-Analysis?
PLoS ONE 8(4): e61505.
labs.genetics.ucla.edu/horvath/CoexpressionNetwork/MetaAnalysis/