

Module preservation statistics

Steve Horvath
University of California, Los Angeles

Module preservation is often an essential step in a network analysis





Construct a network
Rationale: make use of interaction patterns between genes

Identify modules
Rationale: module (pathway) based analysis

Relate modules to external information
Array Information: Clinical data, SNPs, proteomics
Gene Information: gene ontology, EASE, IPA
Rationale: find biologically interesting modules

Study Module Preservation across different data
Rationale:
• Same data: to check robustness of module definition
• Different data: to find interesting modules

Find the key drivers of *interesting* modules
Rationale: experimental validation, therapeutics, biomarkers

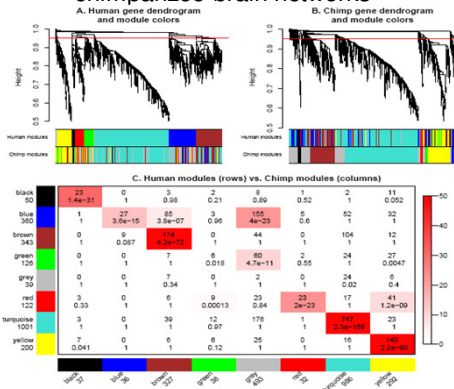





Motivational example: Studying the preservation of human brain co-expression modules in chimpanzee brain expression data.

Modules defined as clusters (branches of a cluster tree)

Data from Oldam et al 2006

Preservation of modules between human and chimpanzee brain networks



Standard cross-tabulation based statistics have severe disadvantages

Disadvantages

1. only applicable for modules defined via a clustering procedure
2. ill suited for making the strong statement that a module is not preserved

We argue that network based approaches are superior when it comes to studying module preservation

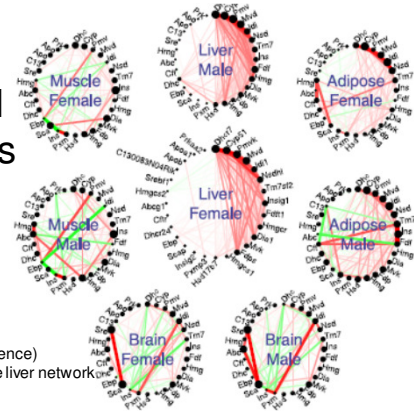
Is my network module preserved and reproducible?

Langfelder et al PloS Comp Biol. 7(1): e1001057.

Broad definition of a module

- Abstract definition of module=subset of nodes in a network.
- Thus, a module forms a sub-network in a larger network
- Example: module (set of genes or proteins) defined using external knowledge: KEGG pathway, GO ontology category
- Example: modules defined as clusters resulting from clustering the nodes in a network
- Module preservation statistics can be used to evaluate whether a given module defined in one data set (reference network) can also be found in another data set (test network)

Network of cholesterol biosynthesis genes



Message:
female liver network (reference)
Looks most similar to male liver network

Question

- How to measure relationships between different networks (e.g. how similar is the female liver network to the male network)?
- Answer: network concepts aka statistics

Connectivity (aka degree)

- Node connectivity = row sum of the adjacency matrix
 - For unweighted networks=number of direct neighbors
 - For weighted networks= sum of connection strengths to other nodes

$$Connectivity_i = k_i = \sum_{j \neq i} a_{ij}$$

$$Scaled\ connectivity = K_i = \frac{k_i}{\max(k)}$$

Density

- Density= mean adjacency
- Highly related to mean connectivity

$$Density = \frac{\sum_i \sum_{j \neq i} a_{ij}}{n(n-1)} = \frac{mean(k)}{n-1}$$

where n is the number of network nodes.

Network concepts to measure relationships between networks

Numerous network concepts can be used to measure the preservation of network connectivity patterns between a reference network and a test network

- E.g. Density in the test set
- $\text{cor.k} = \text{cor}(k^{\text{ref}}, k^{\text{test}})$
- $\text{cor}(A^{\text{ref}}, A^{\text{test}})$

Module preservation in different types of networks

- One can study module preservation in general networks specified by an adjacency matrix, e.g. protein-protein interaction networks.
- However, particularly powerful statistics are available for correlation networks
 - weighted correlation networks are particularly useful for detecting subtle changes in connectivity patterns. But the methods are also applicable to unweighted networks (i.e. graphs)

Network-based module preservation statistics

- Input: module assignment in reference data.
- Adjacency matrices in reference A^{ref} and test data A^{test}
- Network preservation statistics assess preservation of
 - 1. network density: Does the module remain densely connected in the test network?
 - 2. connectivity: Is hub gene status preserved between reference and test networks?
 - 3. separability of modules: Does the module remain distinct in the test data?

Several connectivity preservation statistics

For general networks, i.e. input adjacency matrices

- $\text{cor.kIM} = \text{cor}(kIM^{\text{ref}}, kIM^{\text{test}})$
 - correlation of intramodular connectivity across module nodes
- $\text{cor.ADJ} = \text{cor}(A^{\text{ref}}, A^{\text{test}})$
 - correlation of adjacency across module nodes

For correlation networks, i.e. input sets are variable measurements

- $\text{cor.Cor} = \text{cor}(\text{cor}^{\text{ref}}, \text{cor}^{\text{test}})$
- $\text{cor.kME} = \text{cor}(kME^{\text{ref}}, kME^{\text{test}})$

One can derive relationships among these statistics in case of weighted correlation network

Choosing thresholds for preservation statistics based on permutation test

- For correlation networks, we study 4 density and 4 connectivity preservation statistics that take on values ≤ 1
- Challenge: Thresholds could depend on many factors (number of genes, number of samples, biology, expression platform, etc.)
- Solution: Permutation test. Repeatedly permute the gene labels in the test network to estimate the mean and standard deviation under the null hypothesis of no preservation.

- Next we calculate a Z statistic
$$Z = \frac{\text{observed} - \text{mean}_{\text{permuted}}}{\text{sd}_{\text{permuted}}}$$

Permutation test for estimating Z scores

- For each preservation measure we report the observed value and the permutation Z score to measure significance.

$$Z = \frac{\text{observed} - \text{mean}_{\text{permuted}}}{\text{sd}_{\text{permuted}}}$$

- Each Z score provides answer to "Is the module significantly better than a random sample of genes?"
- Summarize the individual Z scores into a composite measure called Z.summary
- Z.summary < 2 indicates no preservation, 2 < Z.summary < 10 weak to moderate evidence of preservation, Z.summary > 10 strong evidence

Table 1. Overview of module preservation statistics. Details are provided below and in the paper...

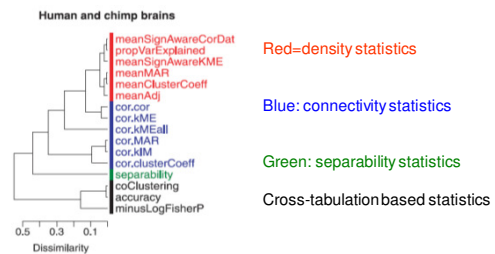
| No. | Preservation Statistic | | Network | Ref. netw. input | | | Test netw. input | | | Used in composite | | |
|-----|-----------------------------|-------|-----------|------------------|-----|------|------------------|-----|------|-------------------|------|-------|
| | Name | Eq. | Type | Lbl | Adj | datX | Lbl | Adj | datX | Zsum. | medR | ZsumA |
| 1 | coClustering | Supp. | Cross-tab | not used | yes | no | no | yes | no | no | no | no |
| 2 | accuracy | Supp. | Cross-tab | not used | yes | no | no | yes | no | no | no | no |
| 3 | -log(p-value) | Supp. | Cross-tab | not used | yes | no | no | yes | no | no | no | no |
| 4 | meanAdj | 8 | Density | general | yes | no | no | no | yes | no | no | yes |
| 5 | meanCoeff | 9 | Density | general | yes | no | no | no | yes | no | no | no |
| 6 | meanMAR | 10 | Density | general | yes | no | no | no | yes | no | no | no |
| 7 | corAdj | 11 | Connect. | general | yes | yes | no | no | yes | no | yes | yes |
| 8 | corKIM | 12 | Connect. | general | yes | yes | no | no | yes | no | yes | yes |
| 9 | corCICoeff | 13 | Connect. | general | yes | yes | no | no | yes | no | no | no |
| 10 | corMAR | 14 | Connect. | general | yes | yes | no | no | yes | no | no | no |
| 11 | separability ^{new} | 27 | Separab. | general | yes | yes | no | no | yes | no | no | no |
| 12 | meanCor | 19 | Den.+Con. | cor | yes | no | yes | no | no | yes | yes | yes |
| 13 | corCor | 20 | Connect. | cor | yes | no | yes | no | no | yes | yes | yes |
| 14 | propVarExpl | 21 | Density | cor | yes | no | yes | no | no | yes | yes | yes |
| 15 | meanKME | 22 | Den.+Con. | cor | yes | no | yes | no | no | yes | yes | yes |
| 16 | corKME | 23 | Connect. | cor | yes | no | yes | no | no | yes | yes | yes |
| 17 | corKMEall | 24 | Connect. | cor | yes | no | yes | no | no | yes | no | no |
| 18 | separability | 28 | Separab. | cor | yes | no | yes | no | no | yes | no | no |
| 19 | Z.summary | 1 | Compos. | cor | yes | yes | yes | no | yes | yes | | |
| 20 | Z.summary | | Compos. | cor | yes | yes | yes | no | yes | yes | | |
| 21 | medianRank | 34 | Compos. | cor | yes | yes | yes | no | yes | yes | | |
| 22 | Z.summary | 35 | Compos. | general | yes | yes | no | no | yes | no | | |

The columns report the names, types, and input of individual preservation statistics (Lbl, module label; Adj, general network adjacency; datX, numeric data from which a correlation network is constructed). The last 3 columns indicate which of the individual statistics are used in the composite summary statistics Z.summary, medianRank, and Z.summary, respectively. The definition of cross-tabulation based statistics can be found in Supplementary Text S1.

Module preservation statistics are often closely related

Message: it makes sense to aggregate the statistics into "composite preservation statistics"

Clustering module preservation statistics based on correlations across modules



Composite statistic in correlation networks based on Z statistics

Permutation test allows one to estimate Z version of each statistic

$$Z_{cor.Cor}^{(q)} = \frac{cor.Cor^{(q)} - E(cor.Cor^{(q)} | null)}{\sqrt{Var(cor.Cor^{(q)} | null)}}$$

Composite connectivity based statistics for correlation networks

$$Z_{connectivity}^{(q)} = median(Z_{cor.Cor}^{(q)}, Z_{cor.kME}^{(q)}, Z_{cor.A}^{(q)}, Z_{cor.kIM}^{(q)})$$

Composite density based statistics for correlation networks

$$Z_{density}^{(q)} = median(Z_{meanCor}^{(q)}, Z_{meanAdj}^{(q)}, Z_{propVarExpl}^{(q)}, Z_{meanKME}^{(q)})$$

Composite statistic of density and connectivity preservation

$$Z_{summary}^{(q)} = \frac{Z_{connectivity}^{(q)} + Z_{density}^{(q)}}{2}$$

Analogously define composite statistic: medianRank

- Based on the ranks of the observed preservation statistics
- Does not require a permutation test
- Very fast calculation
- Typically, it shows no dependence on the module size

Summary preservation

- Network based preservation statistics measure different aspects of module preservation
 - Density-, connectivity-, separability preservation
- Two types of composite statistics: Zsummary and medianRank.
- Composite statistic Zsummary based on a permutation test
 - Advantages: thresholds can be defined, R function also calculates corresponding permutation test p-values
 - Example: Zsummary < 2 indicates that the module is "not" preserved
 - Disadvantages: i) Zsummary is computationally intensive since it is based on a permutation test, ii) often depends on module size
- Composite statistic medianRank
 - Advantages: i) fast computation (no need for permutations), ii) no dependence on module size.
 - Disadvantage: only applicable for ranking modules (i.e. relative preservation)

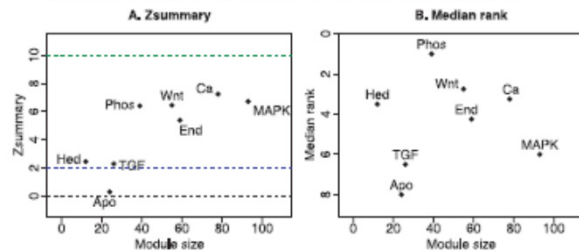
Application: Modules defined as KEGG pathways.

Comparison of human brain (reference) versus chimp brain (test) gene expression data.

Connectivity patterns (adjacency matrix) is defined as signed weighted co-expression network.

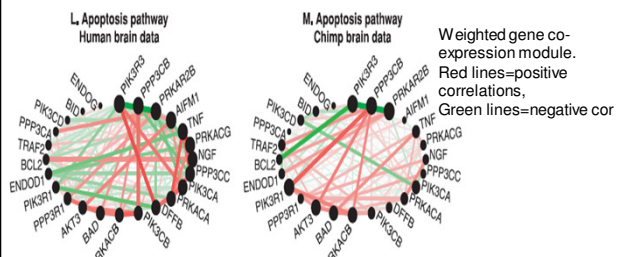
Preservation of KEGG pathways measured using the composite preservation statistics Zsummary and medianRank

- Humans versus chimp brain co-expression modules



Apoptosis module is least preserved
according to both composite preservation statistics

Visually inspect connectivity patterns of the apoptosis module in humans and chimpanzees



Note that the connectivity patterns look very different.
Preservation statistics are ideally suited to measure differences
in connectivity preservation

Literature validation:

Neuron apoptosis is known to differ
between humans and chimpanzees

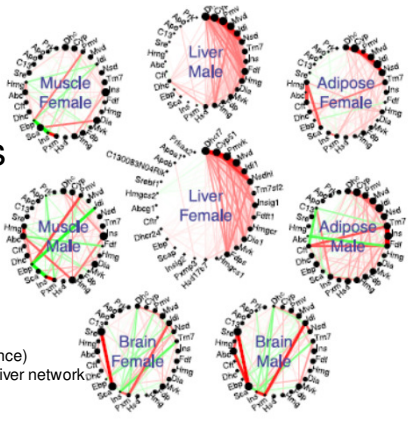
- It has been hypothesized that natural selection for increased cognitive ability in humans led to a reduced level of neuron apoptosis in the human brain:
 - Arora et al (2009) Did natural selection for increased cognitive ability in humans lead to an elevated risk of cancer? *Med Hypotheses* 73: 453–456.
- Chimpanzee tumors are extremely rare and biologically different from human cancers
- A scan for positively selected genes in the genomes of humans and chimpanzees found that a large number of genes involved in apoptosis show strong evidence for positive selection (Nielsen et al 2005 *PloS Biol*).

Application: Studying the preservation of a female mouse liver module in different tissue/gender combinations.

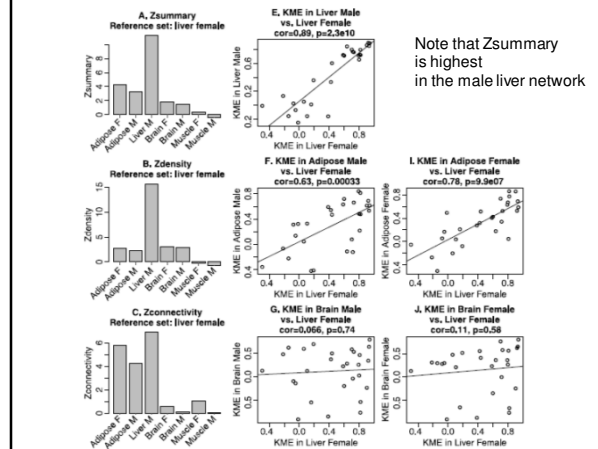
Module: genes of cholesterol biosynthesis pathway
Network: signed weighted co-expression network
Reference set: female mouse liver
Test sets: other tissue/gender combinations

Data provided by Jake Lusis

Network of cholesterol biosynthesis genes



Message:
female liver network (reference)
Looks most similar to male liver network



Publicly available microarray data
from
lung adenocarcinoma patients

References of the array data sets

- Shedden et al (2008) Nat Med. 2008 Aug;14(8):822-7
- Tomida et al (2009) J Clin Oncol 2009 Jun 10;27(17):2793-9
- Bild et al (2006) Nature 2006 Jan 19;439(7074):353-7
- Takeuchi et al (2006) J Clin Oncol 2006 Apr 10;24(11):1679-88
- Roepman et al (2009) Clin Cancer Res. 2009 Jan 1;15(1):284-90

Array platforms

5 Affymetrix data sets

- Affy 133 A – Shedden et al (HLM, Mich, MSKCC, DFCI)
- Affy 133 plus 2 – Bild et al

3 Agilent platforms:

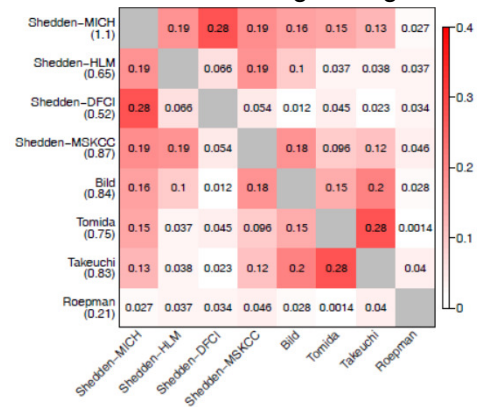
- 21.6K custom array – Takeuchi et al
- Whole Human Genome Microarray 4x44K – Tomida et al
- Whole Human Genome Oligo Microarray G4112A – Roepman et al

Standard marginal analysis
for relating genes to survival time

(Prognostic) Gene Significance

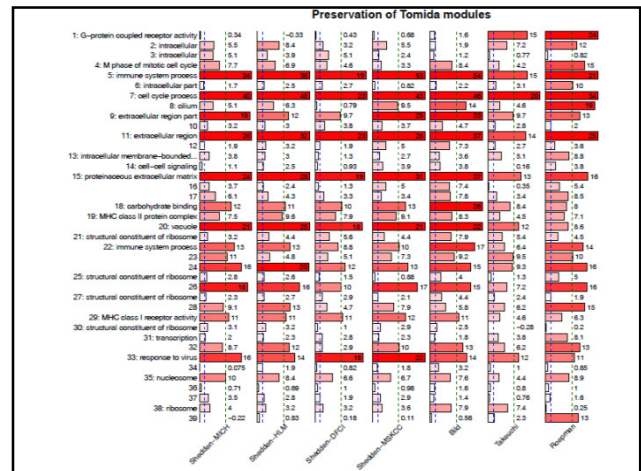
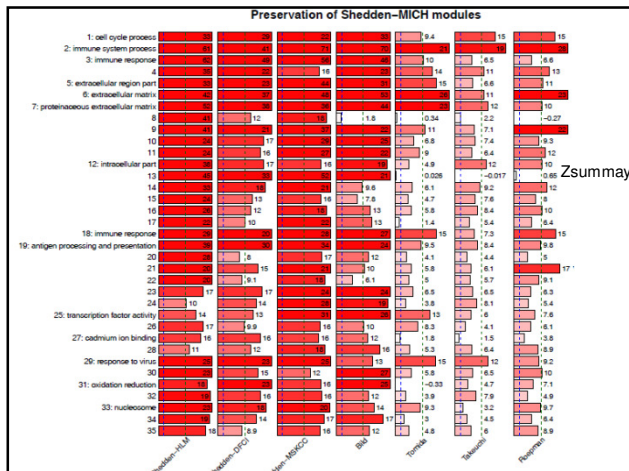
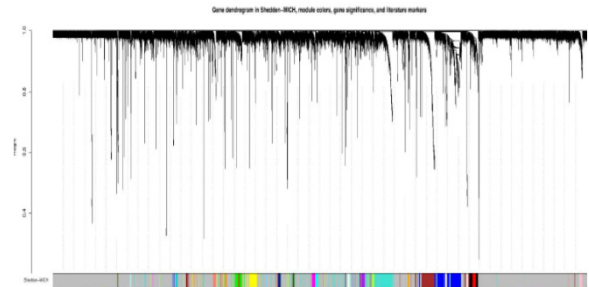
- Roughly speaking: the correlation between gene expression and survival time.
- More accurately: relation to hazard of death (Cox regression model)

Weak relations between gene significances

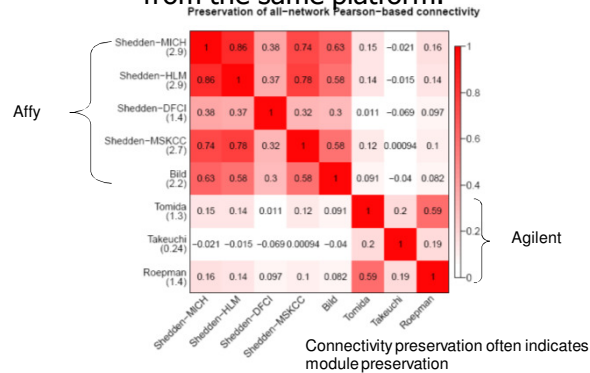


Gene co-expression module preservation

Modules found in the Shedden Michigan data set;



Adenocarcinoma: Network connectivity is correlated for data from the same platform.



Consensus module analysis

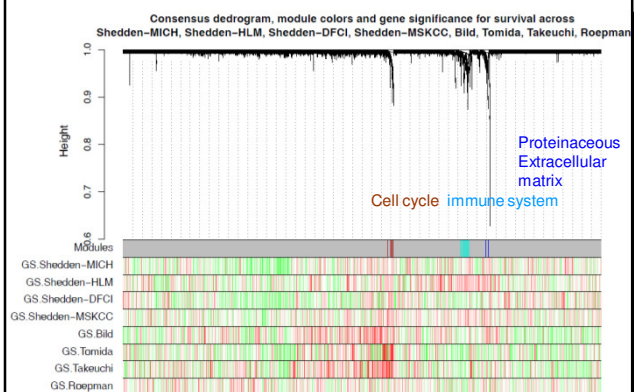
Steps for defining “consensus” modules that are shared across many networks

- Calibrate individual networks so that they become comparable
 - Often easier for weighted networks
- Define consensus network using quantile

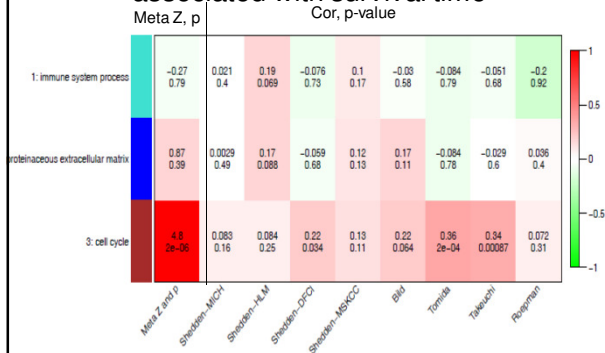
$$A_{\text{cons}}_{ij} = p\text{quantile}(A^{(1)}_{ij}, A^{(2)}_{ij}, \dots, \text{prob} = .25)$$

- Define consensus dissimilarity based on consensus network
- Define modules as clusters
- Use WGCNAR function `blockwiseConsensusModules` or `consensusDissTOMandTree`

Consensus modules based on 8 adeno data sets



As expected, the cell cycle module eigengene is significantly ($p=2E-6$) associated with survival time



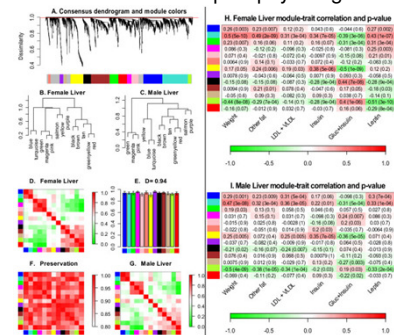
Advantages of soft thresholding with the power function

1. Robustness: Network results are highly robust with respect to the choice of the power beta (Zhang et al 2005)
2. Calibrating different networks becomes straightforward, which facilitates consensus module analysis
3. Math reason: Geometric Interpretation of Gene Co-Expression Network Analysis. PloS Computational Biology. 4(8): e1000117
4. Module preservation statistics are particularly sensitive for measuring connectivity preservation in weighted networks

Another application:
Consensus modules in male and female liver tissue

$$A.cons_{ij} = \min(A^{(female)}_{ij}, A^{(male)}_{ij})$$

Consensus eigengene networks in male and female mouse liver data and their relationship to physiological traits



Langfelder P et al (2007) Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology 2007

Implementation and R software tutorials, WGCNA R library

- General information on weighted correlation networks
- Google search
 - “WGCNA”
 - “weighted gene co-expression network”
- R function `modulePreservation` is part of WGCNA package
- Tutorials: preservation between human and chimp brains

www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/ModulePreservation

Acknowledgement

Students and Postdocs:

- **Peter Langfelder** first author on many related articles
- Jason Aten, Chaochao (Ricky) Cai, Jun Dong, Tova Fuller, Ai Li, Wen Lin, Michael Mason, Jeremy Miller, Mike Oldham, Chris Plaisier, Anja Presson, Lin Song, Kellen Winden, Yafeng Zhang, Andy Yip, Bin Zhang
- Colleagues/Collaborators
- Cancer: Paul Mischel, Stan Nelson
- Neuroscience: Dan Geschwind, Giovanni Coppola, Roel Ophoff
- Mouse: Jake Lusis, Tom Drake