Work Review
2011.3- 2013.3
Beibei Xin
Research Experiences

• 2011.3-2011.5      Internship
  Define A Subspace-based Similarity Measurement in Microarray Data

• 2012.6-2012.8      Lab project
  Detecting piRNAs During Zebrafish Embryo Development: RNA-seq Analysis

• 2012.9-2013.3      Lab project
  High-resolution and Genome-wide Detection of Imprinting in Zebrafish
Define A Subspace-based Similarity Measurement in Microarray Data
2011.3-2011.5

Bing Zhang et al., 2005
Weighted Gene Co-expression Network Analysis Pipeline

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 in interesting modules
Tools: intramodular connectivity, causality testing
Rationale: experimental validation, therapeutics, biomarkers

Bing Zhang et al., 2005
Steps for Constructing WGCN

- Gene list \( L \) to be analyzed: differential expressed gene list (tumor vs. normal)
- Determine parameters: \( \beta, a \)
- Calculate adjacency matrix

\[
A = (a_{ij})_{i,j=1\ldots N} = \{(\text{sim}(a_i, a_j))^\beta\}_{i,j=1\ldots N}
\]

- Calculate distance matrix

\[
\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}
\]

where \( l_{ij} = \sum_u a_{iu}a_{uj} \), and \( k_i = \sum_u a_{iu} \)

\[
D = (1 - \omega_{ij})_{i,j=1\ldots N}
\]
A good similarity measurement is ...

- Three mathematical properties: between 0 and 1, symmetry principle, triangle inequality
- Free from effects of outliers
- Robust

Commonly used: Pearson correlation

Calculated in whole space, affected by detecting error
Procedure to find subspace S

Raw microarray

Mesh every sample data into N/a intervals

Order matrix

Find most frequent order for every gene

Representative order

\[
\begin{align*}
gen1 &: a_{11}, a_{12}...a_{1d} \\
gene2 &: a_{21}, a_{22}...a_{2d} \\
&... \\
geneN &: a_{N1}, a_{N2}...a_{Nd}
\end{align*}
\]

\[
\begin{align*}
gen1 &: \text{order}^{(1)}(a_1), \text{order}^{(2)}(a_1)...\text{order}^{(d)}(a_1) \\
gene2 &: \text{order}^{(1)}(a_2), \text{order}^{(2)}(a_2)...\text{order}^{(d)}(a_2) \\
&... \\
geneN &: \text{order}^{(1)}(a_N), \text{order}^{(2)}(a_N)...\text{order}^{(d)}(a_N)
\end{align*}
\]

\[
\begin{align*}
gen1 &: \text{order}(a_1) \\
gene2 &: \text{order}(a_2) \\
&... \\
geneN &: \text{order}(a_N)
\end{align*}
\]
Steps:
1. Sort microarray data in ascending order at every sample
2. Divide every sample into N/a intervals and give them orders
3. Find most frequent order for N genes
4. Determination of subspace $S$ for gene $x$ and $y$

$$S(x) = \{ \eta_i \mid \text{order}^{(\eta_i)}(x) = \text{order}(x), 1 \leq \eta_i \leq d \}$$

$$S = S(x) \cap S(y)$$
Sub-space Based Similarity Function

The measure is

$$subsim^{(S)}(x, y) = \frac{d(S)}{d} \sum_{i=1}^{d(S)} sim^{(i)}(x, y)$$

Where $d(s)<d$ and

$$sim^{(i)}(x, y) = \begin{cases} 
1 - \frac{|x_i - y_i|}{UR_i,order(x) - LR_i,order(x)}, & order(x) = order(y); \\
1 - \frac{1}{2} \cdot \frac{UR_i,order(x) - x_i}{UR_i,order(x) - LR_i,order(x)} - \frac{1}{2} \cdot \frac{y_i - LR_i,order(y)}{UR_i,order(y) - LR_i,order(y)}, & order(x) < order(y); \\
1 - \frac{1}{2} \cdot \frac{x_i - LR_i,order(x)}{UR_i,order(x) - LR_i,order(x)} - \frac{1}{2} \cdot \frac{UR_i,order(y) - y_i}{UR_i,order(y) - LR_i,order(y)}, & order(x) > order(y). 
\end{cases}$$
Quality proof

- Sub-space based similarity is between 0 and 1, satisfies symmetry principle
- Defined in subspace, it removes outlier sample data
- The following gradual theorem is a criteria to obtain a robust measurement.

Assume $X$ is a random variable, $\frac{D_{\text{max}}^d - D_{\text{min}}^d}{D_{\text{min}}^d} \to 0$ when $\lim_{d \to \infty} \var\left( \frac{X^d}{E(|X^d|)} \right) = 0$, 

$$ \frac{D_{\text{max}}^d - D_{\text{min}}^d}{D_{\text{min}}^d} \to 0 \quad \text{when} \quad \lim_{d \to \infty} \var\left( \frac{X^d}{E(|X^d|)} \right) = 0, $$
Two criterions to determine $\beta$ and $a$

- **Scale-free topology**: Indeed, recent studies have demonstrated that the probability that a substrate can react with $k$ other substrates [the degree distribution $P(k)$ of a metabolic network] decays as a power law $P(k) \sim c k^{-\gamma}$ with $\gamma \approx 2.2$ in all organisms.

- **Ensure a robust measure**: for $a$, the less, the better.

Note: when dealing with concrete case, these two criterions can be used to compromise a set of proper parameter values.
Case study 1: clear cell renal cell carcinoma

A) Sub-space
B) Pearson
C) $k=-2.27$
D) $k=-1.11$
Case study 2: breast cancer

Functions with Bonferroni value <0.05 in every module are listed out

Gene Category
- mitotic cell cycle
- cell proliferation
- cell cycle
- antigen presentation
- cellular process
- antigen processing

Module 1:
- mitotic cell cycle
- cell proliferation
- cell cycle
- antigen presentation
- cellular process
- antigen processing

Module 2:
- response to pest/pathogen/parasite
- immune response
- response to external stimulus
- response to wounding
- response to biotic stimulus

Module 3:
- defense response
- cell communication
- response to stress
- cell surface receptor linked signal transduction
- signal transduction
- cellular process
- inflammatory response
- lipid metabolism
- innate immune response
- carbohydrate metabolism
- alcohol metabolism
- response to chemical substance
- carboxylic acid metabolism
- organic acid metabolism
- taxis
- chemotaxis
- cell-cell signaling

Module 4:
- immune response
- defense response
- response to biotic stimulus
- inflammatory response
- innate immune response
- response to pest/pathogen/parasite
- response to stress
Case study 3: lung cancer
Detecting piRNAs During Zebrafish Embryo Development: RNA-seq Analysis

2012.6-2012.8
Intuitions to find the molecular mechanism:

- small RNA from germ-line cells mediates DNA methylation reprogramming to guide epigenetic inheritance in plant (Calarco et al., 2012)

- Piwi-interacting RNA (piRNA) mediates DNA methylation of transposon elements during germ cell maturation in mouse (Aravin et al., 2007, 2008)

- piRNAs is abundant in zebrafish sperm (Houwing et al., 2007)
RNA-seq Data for sperm, egg and 8 stages

1-cell (0.2 h)
16-cell (1.5 h)
32-cell (1.75 h)
64-cell (2 h)
128-cell (2.25 h)
1k-cell (3 h)
sphere (4 h)
germ ring (5.7 h)
Detecting piRNA Pipeline

All small RNAs reads

After clipping the adapter sequences, the remaining sequence is 16-32 nt?

Yes

Mapping to genome?

No

Is known ncRNAs?

Yes

Mapping to transcriptom?

24-32 nt?

Yes

piRNA reads

No
piRNA Biogenesis -- Pingpong Model

Alexi A. Aravin et al. 2007
High-resolution and Genome-wide Detection of Imprinting in Zebrafish

2012.9-2013.3
Background

Imprinting

Wolf Reik et al. 2001
Background

Found in human, eutherian mammals, marsupials, and flowering plants
- 45 imprinted genes in mouse
- Imprinting disorder leads to Prader-Willi syndrome in human
- No genome-wide imprinting evidence in Zebrafish

Characteristics of imprinting genes
- around 80% are physically linked in clusters with other imprinted genes (ICs)
- rich in CpG islands
- clustered, direct repeats are common near to or within the CpG islands
Background

Figure A: sites at least four consecutive oocyte-specific unmethylated CpG sites
Figure B: sites at least four consecutive sperm-specific unmethylated CpG sites
Two approaches to detect imprinting genes:
Experimental strategy

Use exited high-resolution data of 4 stages to find potential Allele-specific Methylated Regions

Validation (Cross breeding)

DNA methylation screen sperm, egg, 1k-cell, germ ring
Methylation levels follow dynamic pattern

Gene expression screen 1k-cell, germ ring
Only one parent-of-origin allele has significant gene expression
28 Potential AMRs

Pair-end reads of whole-genome Bisulfite sequencing from sperm, egg, 1kcell and germring

amrfinder software (Fang Fang, et al.) finds AMRs for every stage

AMRs overlap of four stages with methylation level
Sperm<0.25 or >0.75
Egg >0.75 or <0.25
1kcell between 0.4 to 0.6
Germring between 0.4 to 0.6

<table>
<thead>
<tr>
<th>chromosome</th>
<th>StartPos</th>
<th>EndPos</th>
<th>#CpG</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>59214836</td>
<td>59214856</td>
<td>5</td>
</tr>
<tr>
<td>chr2</td>
<td>17381222</td>
<td>17381233</td>
<td>3</td>
</tr>
<tr>
<td>chr4</td>
<td>1751478</td>
<td>1751522</td>
<td>3</td>
</tr>
<tr>
<td>chr5</td>
<td>71761203</td>
<td>71761261</td>
<td>7</td>
</tr>
<tr>
<td>chr6</td>
<td>45764515</td>
<td>45764679</td>
<td>10</td>
</tr>
<tr>
<td>chr8</td>
<td>29554684</td>
<td>29554840</td>
<td>7</td>
</tr>
<tr>
<td>chr9</td>
<td>6755024</td>
<td>6755048</td>
<td>2</td>
</tr>
<tr>
<td>chr10</td>
<td>15788714</td>
<td>15788745</td>
<td>3</td>
</tr>
<tr>
<td>chr10</td>
<td>37953104</td>
<td>37953111</td>
<td>2</td>
</tr>
<tr>
<td>chr13</td>
<td>22074836</td>
<td>22074864</td>
<td>3</td>
</tr>
<tr>
<td>chr15</td>
<td>30487764</td>
<td>30487831</td>
<td>7</td>
</tr>
<tr>
<td>chr17</td>
<td>6080339</td>
<td>6080345</td>
<td>3</td>
</tr>
<tr>
<td>chr17</td>
<td>15713328</td>
<td>15713397</td>
<td>6</td>
</tr>
<tr>
<td>chr17</td>
<td>22341958</td>
<td>22341975</td>
<td>3</td>
</tr>
<tr>
<td>chr18</td>
<td>2337731</td>
<td>23377823</td>
<td>5</td>
</tr>
<tr>
<td>chr19</td>
<td>27171711</td>
<td>27171764</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>29670691</td>
<td>29670717</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>43272496</td>
<td>43272562</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>44027674</td>
<td>44027701</td>
<td>3</td>
</tr>
<tr>
<td>chr20</td>
<td>16704363</td>
<td>16704379</td>
<td>2</td>
</tr>
<tr>
<td>chr21</td>
<td>32052512</td>
<td>32052528</td>
<td>3</td>
</tr>
<tr>
<td>chr22</td>
<td>32671940</td>
<td>32671959</td>
<td>2</td>
</tr>
<tr>
<td>chr23</td>
<td>35234004</td>
<td>35234028</td>
<td>3</td>
</tr>
<tr>
<td>chr24</td>
<td>22543932</td>
<td>22544075</td>
<td>9</td>
</tr>
<tr>
<td>chr24</td>
<td>24286664</td>
<td>24286690</td>
<td>4</td>
</tr>
<tr>
<td>chr24</td>
<td>25671897</td>
<td>25671912</td>
<td>3</td>
</tr>
<tr>
<td>chr25</td>
<td>10839805</td>
<td>10839805</td>
<td>3</td>
</tr>
<tr>
<td>chr25</td>
<td>29376890</td>
<td>29376910</td>
<td>3</td>
</tr>
</tbody>
</table>
Cross breeding

Long fin  Tub

Sperm  egg  1kcell  germring

Genomic DNA

AMRs enrichment by PCR

Illumina library
<table>
<thead>
<tr>
<th>chromosome</th>
<th>StartPos</th>
<th>EndPos</th>
<th>#CpG</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>59214836</td>
<td>59214856</td>
<td>5</td>
</tr>
<tr>
<td>chr2</td>
<td>17381222</td>
<td>17381233</td>
<td>3</td>
</tr>
<tr>
<td>chr4</td>
<td>1751478</td>
<td>1751522</td>
<td>3</td>
</tr>
<tr>
<td>chr5</td>
<td>71761203</td>
<td>71761261</td>
<td>7</td>
</tr>
<tr>
<td>chr6</td>
<td>45764515</td>
<td>45764679</td>
<td>10</td>
</tr>
<tr>
<td>chr8</td>
<td>29554684</td>
<td>29554840</td>
<td>7</td>
</tr>
<tr>
<td>chr9</td>
<td>6755024</td>
<td>6755048</td>
<td>2</td>
</tr>
<tr>
<td>chr10</td>
<td>15788714</td>
<td>15788745</td>
<td>3</td>
</tr>
<tr>
<td>chr10</td>
<td>37953104</td>
<td>37953111</td>
<td>2</td>
</tr>
<tr>
<td>chr13</td>
<td>22074836</td>
<td>22074864</td>
<td>3</td>
</tr>
<tr>
<td>chr15</td>
<td>30487764</td>
<td>30487831</td>
<td>7</td>
</tr>
<tr>
<td>chr17</td>
<td>6080339</td>
<td>6080345</td>
<td>3</td>
</tr>
<tr>
<td>chr17</td>
<td>15713328</td>
<td>15713397</td>
<td>6</td>
</tr>
<tr>
<td>chr17</td>
<td>22341958</td>
<td>22341975</td>
<td>3</td>
</tr>
<tr>
<td>chr18</td>
<td>23377731</td>
<td>23377823</td>
<td>5</td>
</tr>
<tr>
<td>chr19</td>
<td>27171711</td>
<td>27171764</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>29670691</td>
<td>29670717</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>43272496</td>
<td>43272562</td>
<td>4</td>
</tr>
<tr>
<td>chr19</td>
<td>44027674</td>
<td>44027701</td>
<td>3</td>
</tr>
<tr>
<td>chr20</td>
<td>16704363</td>
<td>16704379</td>
<td>2</td>
</tr>
<tr>
<td>chr21</td>
<td>32052512</td>
<td>32052528</td>
<td>3</td>
</tr>
<tr>
<td>chr22</td>
<td>32671940</td>
<td>32671959</td>
<td>2</td>
</tr>
<tr>
<td>chr23</td>
<td>35234004</td>
<td>35234028</td>
<td>3</td>
</tr>
<tr>
<td>chr24</td>
<td>22543932</td>
<td>22544075</td>
<td>9</td>
</tr>
<tr>
<td>chr24</td>
<td>24286664</td>
<td>24286690</td>
<td>4</td>
</tr>
<tr>
<td>chr24</td>
<td>25671897</td>
<td>25671912</td>
<td>3</td>
</tr>
<tr>
<td>chr25</td>
<td>10839785</td>
<td>10839805</td>
<td>3</td>
</tr>
<tr>
<td>chr25</td>
<td>29376890</td>
<td>29376910</td>
<td>3</td>
</tr>
</tbody>
</table>
Many thanks to Professor Liu Jiang’s Lab members

Thank you very much!