Guided Tutorial:

Using GSEA as an analytical tool for molecular profiling
Getting Started:
Create a GSEA workstation that works for you

When would you use GSEA?

What do you need to get started?

Basic workstation with Java
Expression files converted to .gct (GenePattern)
Phenotype labels (any text editor)
Gene Sets (public repository or custom sets)

Desktop/Laptop running Mac OS X, Windows, or Linux:

Java 6 or 7 (Desktop GUI or command line)

R

GenePattern (Web, Desktop, or Server)
Getting Started:
Working with GSEA files

1. Sources of gene expression data (.gct files):
   Microarrays (.cel files)
   RNA-Seq (.fpkm_tracking file)

2. Phenotype labels (.cls files)

3. Gene Sets (MSigDB or custom):
   Positional
   Curated
   Pathway (KEGG, Biocarta, etc.)
### Getting Started:
#### GCT file column structure

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 1</td>
<td>Row identifiers. Typically probe set ids or clone ids. These must be UNIQUE</td>
</tr>
<tr>
<td>Column 2</td>
<td>Row descriptions. Ignored by the program – can be dummy values (e.g. “na”)</td>
</tr>
<tr>
<td>Each column</td>
<td>Each column contains expression values from 1 sample. Missing values are allowed (leave empty).</td>
</tr>
</tbody>
</table>

#### Example GCT file content:

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Description</th>
<th>Untreated FPKM</th>
<th>Cisplatin FPKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mcm4</td>
<td>chr16:15623989-15637493</td>
<td>16.5908</td>
<td>33.9435</td>
</tr>
<tr>
<td>Mcm5</td>
<td>chr8:77633426-77652338</td>
<td>5.97553</td>
<td>16.5161</td>
</tr>
<tr>
<td>Mcm6</td>
<td>chr1:130228167-130256233</td>
<td>12.7255</td>
<td>28.2007</td>
</tr>
<tr>
<td>Mcm7</td>
<td>chr5:138605816-138613090</td>
<td>25.7281</td>
<td>49.4802</td>
</tr>
</tbody>
</table>

If editing in Excel, make sure to save your data as “tab delimited text.”
### Getting Started: Converting from microarray .cel expression to .gct file

**ExpressionFileCreator**

<table>
<thead>
<tr>
<th>Required Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input file*</td>
<td>Choose File</td>
</tr>
<tr>
<td>Method*</td>
<td>RMA</td>
</tr>
<tr>
<td>Quantile normalization</td>
<td>yes</td>
</tr>
<tr>
<td>Background correct</td>
<td>yes</td>
</tr>
<tr>
<td>Compute present absent calls</td>
<td>no (create gct file)</td>
</tr>
<tr>
<td>Normalization method</td>
<td>Median scaling</td>
</tr>
<tr>
<td>Value to scale to</td>
<td>Median/mean scaling only</td>
</tr>
<tr>
<td>CMM file</td>
<td>Choose File</td>
</tr>
<tr>
<td>Annotate probes*</td>
<td>yes</td>
</tr>
<tr>
<td>CDF file</td>
<td>Choose File</td>
</tr>
<tr>
<td>Output file*</td>
<td><code>&lt;input_file_basename&gt;</code></td>
</tr>
</tbody>
</table>

*Note: The `input_file` is a single .cel file to be uploaded.*
Getting Started:
Converting from RNA-Seq expression genes.fpkm_tracking to .gct file

<table>
<thead>
<tr>
<th>Fpkm_trackingToGct</th>
<th>version 3</th>
</tr>
</thead>
</table>

* required field

**input file**
- Choose file: No file chosen
- Select a single file under 2GB to upload.
- Specify URL: Basic Upload

**row labels**
- Tracking ID

**row descriptions**
- Locus

**filter rows**
- no

**output columns**
- FPKM

**output prefix**
- `<input.file.basename>`

View code to call Fpkm_trackingToGct:
Getting Started:
Real World Example

GCT file structure:

```
#1.2
23230  2

<table>
<thead>
<tr>
<th>gene_short_name</th>
<th>Description</th>
<th>Untreated</th>
<th>Cisplatin</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneA</td>
<td>na</td>
<td>16.5908</td>
<td>33.9435</td>
</tr>
<tr>
<td>GeneB</td>
<td>na</td>
<td>5.97553</td>
<td>16.5161</td>
</tr>
<tr>
<td>GeneC</td>
<td>na</td>
<td>12.7255</td>
<td>28.2007</td>
</tr>
<tr>
<td>GeneD</td>
<td>na</td>
<td>25.7281</td>
<td>49.4802</td>
</tr>
</tbody>
</table>
```
Getting Started: Creating a .cls file to label phenotypes

Example of a 3 class cls file

```
3 1
# KRAS_MUT WT MYC_MUT
KMUT KMUT KMUT WT WT WT myc myc myc
```
### GCT file structure:

```
#1.2
23230  2
gene_short_name  Description   Untreated   Cisplatin
GeneA    na  16.5908  33.9435
GeneB    na  5.97553  16.5161
GeneC    na  12.7255  28.2007
GeneD    na  25.7281  49.4802
```

### CLS file structure:

```
2 2 1
# Untreated Cisplatin
Untreated Cisplatin
```
Running GSEA:
Analyzing cancer cell lines for p53 targets

Download and load the p53 datasets from the GSEA website:
http://www.broadinstitute.org/gsea/datasets.jsp

P53_hgu95av2 (50 p53 MUT or WT cell lines analyzed on the HG_U95Av2 (Affy) Human Genome U95 GeneChip:

<table>
<thead>
<tr>
<th>#</th>
<th>NAME</th>
<th>Description</th>
<th>786-0</th>
<th>BT-549</th>
<th>...</th>
<th>UACC-62</th>
<th>UO-31</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>100_g_at</td>
<td>na</td>
<td>215.37</td>
<td>132.94</td>
<td></td>
<td>451.01</td>
<td>186.68</td>
</tr>
<tr>
<td>12625</td>
<td>1000_at</td>
<td>na</td>
<td>328.68</td>
<td>234.31</td>
<td></td>
<td>381.41</td>
<td>285.72</td>
</tr>
<tr>
<td>1001_at</td>
<td>na</td>
<td></td>
<td>39.64</td>
<td>8.84</td>
<td></td>
<td>40.52</td>
<td>31.88</td>
</tr>
</tbody>
</table>

P53.cls (MUT vs. WT):
50 2 1
#MUT WT
MUT MUT ... WT WT

Chip: HG_U95Av2.chip
Running GSEA: Analyzing cancer cell lines for p53 targets
Running GSEA:
Analyzing cancer cell lines for p53 targets
GSEA Reports:
Analyzing cancer cell lines for p53 targets

Enrichment in phenotype: WT (17 samples)
- 1603 / 3437 gene sets are upregulated in phenotype WT
- 11 gene sets are significant at FDR < 25%
- 37 gene sets are significantly enriched at nominal p-value < 1%
- 117 gene sets are significantly enriched at nominal p-value < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: MUT (33 samples)
- 1834 / 3437 gene sets are upregulated in phenotype MUT
- 1 gene set is significantly enriched at FDR < 25%
- 8 gene sets are significantly enriched at nominal p-value < 1%
- 53 gene sets are significantly enriched at nominal p-value < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details
- The dataset has 12625 native features
- After collapsing features into gene symbols, there are: 9096 genes

Gene set details
- Gene set size filters (min=15, max=500) resulted in filtering out 1413 / 4850 gene sets
- The remaining 3437 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)
GSEA Reports:
Analyzing cancer cell lines for p53 targets
### GSEA Reports:
Analyzing cancer cell lines for p53 targets

#### Enrichment plot: BIOCARTA_P53_PATHWAY

<table>
<thead>
<tr>
<th>Upregulated in class</th>
<th>WT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeneSet</td>
<td>BIOCARTA_P53_PATHWAY</td>
</tr>
<tr>
<td>Enrichment Score (ES)</td>
<td>0.7464528</td>
</tr>
<tr>
<td>Normalized Enrichment Score (NES)</td>
<td>2.1340156</td>
</tr>
<tr>
<td>Nominal p-value</td>
<td>0.0</td>
</tr>
<tr>
<td>FDR q-value</td>
<td>0.021385321</td>
</tr>
<tr>
<td>FWER p-Value</td>
<td>0.055</td>
</tr>
</tbody>
</table>
Leading Edge Analysis:
Analyzing cancer cell lines for p53 targets

CSEA v2.0.12 (Gene set enrichment analysis -- Broad Institute)

Select a GSEA result from the application cache
[ OR ] Locate a GSEA result folder from the file system

GSEA Results

Filter Gene Sets

positive phenotype: na pos negative phenotype: MUT

3437 out of 3437 gene sets

Gene Set: Size ES NES NOM p-val FDR q- FWER p-val Rank at Max Leading Ed...

AMUNDSON_DNA_DAMAGE_RESPONSE_TPS3 15 0.852 2.266 0 0.003 0.006 298 tags=60%
MANNS_RESPONSE_TO_AMIFOSTINE_UP 20 0.684 2.287 0 0.006 0.005 193 tags=30%
WARTERS_IR_RESPONSE_SCY 21 0.785 2.187 0 0.023 0.053 152 tags=31%
BIOCARTA_P53_PATHWAY 15 0.747 2.134 0 0.023 0.053 152 tags=31%
WARTERS_RESPONSE_TO_IRSKIN 36 0.63 2.109 0 0.028 0.088 1,130 tags=44%
LASTOWSKA_NEUROBLASTOMA_COPY_NUMBER_DN 354 -0.503 -2.131 0 0.042 0.038 1,181 tags=49%
WILENSKY_RESPONSE_TO_DAPLADIB 23 0.649 2.037 0 0.070 0.227 1,941 tags=57%
INGA_TPS3_TARGETS 15 0.706 2.024 0 0.079 0.256 52 tags=22%
ODONNE_METASTASIS_UP 52 0.541 1.958 0 0.176 0.515 1,759 tags=46%
REACTOME_AMINE_LIGAND_BINDING_RECEPTORS 21 0.706 1.933 0.002 0.217 0.61 977 tags=52%
BIOCARTA_HSP27_PATHWAY 15 0.687 1.904 0.002 0.25 0.716 723 tags=40%
REACTOME_PEP_TIDE_BINDING_RECEPTORS 121 0.476 1.912 0 0.25 0.685 3,040 tags=60%
MATZUR_SPERMATOGONIA 17 0.636 1.891 0.002 0.271 0.761 562 tags=24%
REACTOME_PHASE_II_CONJUGATION 39 0.301 1.871 0.004 0.314 0.824 3,172 tags=56%
NIKOLSKY_BREAST_CANCER_16Q24_AMPLICON 23 0.667 1.86 0 0.33 0.853 2,352 tags=70%
KRELEY_RESPONSE_TO_CISPLATIN_UP 33 0.586 1.834 0 0.382 0.913 298 tags=27%
REACTOME_CLASS_A1_RHODOPSINLIKE_RECEPTORS 186 0.466 1.818 0.002 0.421 0.953 2,164 tags=44%
BIOCARTA_CK1_PATHWAY 16 0.569 1.798 0.006 0.483 0.961 1,239 tags=50%
LIAN_UPA_TARGETS_3M 37 0.345 1.763 0.002 0.634 0.986 2,061 tags=57%
VIJHAS_NOTCH1_TARGETS_DN 15 0.898 1.758 0 0.834 0.987 1,556 tags=60%
BIOCARTA_INFLAM_PATHWAY 28 0.585 1.751 0.019 0.845 0.991 2,415 tags=57%
TAYLOR_METHYLATED_IN_ACUTELYMPHOBLASTICLEUKEMIA 37 -0.549 -1.931 0 0.647 0.512 2,073 tags=62%
NIELSEN_GST_AND_SYNOVAL_SARCOMA_DN 19 0.565 1.729 0.009 0.682 0.996 1,532 tags=37%
REACTOME_GP_CER_DOWNSTREAM_SIGNALING 328 0.397 1.706 0.004 0.684 0.997 2,414 tags=43%
ESBAUER_TARGETS_OF_PAX3_FOXO1_FUSION_DN 30 0.604 1.701 0.002 0.686 0.997 2,384 tags=50%
KANNAN_TPS3_TARGETS_UP 52 0.454 1.684 0.002 0.699 0.998 371 tags=17%
REACTOME_DAG_AND_IP3_SIGNALING 29 0.544 1.709 0.004 0.69 0.997 2,157 tags=55%
ONGUSHI_TPS3_TARGETS 28 0.5 1.696 0.007 0.693 0.998 23 tags=11%

For 8 selected gene sets: Run leading edge analysis Build HTML Report
Leading Edge Analysis: Analyzing cancer cell lines for p53 targets
Leading Edge Analysis: Analyzing cancer cell lines for p53 targets
Browsing MSigDB:
Analyzing cancer cell lines for p53 targets

Description

p53 is a transcription factor whose activity is regulated by phosphorylation. The function is p53 is to keep the cell from progressing through the cell cycle if there is damage to DNA present. It may do this in multiple ways from holding the cell at a checkpoint until repairs can be made to causing the cell to enter apoptosis if the damage cannot be repaired. The critical role of p53 is evidenced by the fact that it is mutated in a very large fraction of tumors from nearly all sources.

Genes in the set

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F1</td>
<td>E2F transcription factor 1</td>
</tr>
<tr>
<td>TP53</td>
<td>tumor protein p53 (Li-Fraumeni,...</td>
</tr>
<tr>
<td>RB1</td>
<td>retinoblastoma 1 (including ost...</td>
</tr>
<tr>
<td>CDK4</td>
<td>cyclin-dependent kinase 4</td>
</tr>
<tr>
<td>TIMP3</td>
<td>TIMP metalloproteinase inhibitor</td>
</tr>
<tr>
<td>ATM</td>
<td>ataxia telangiectasia mutated (...</td>
</tr>
<tr>
<td>CDK2</td>
<td>cyclin-dependent kinase 2</td>
</tr>
<tr>
<td>CCNE1</td>
<td>cyclin E1</td>
</tr>
<tr>
<td>CCND1</td>
<td>cyclin D1</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>cyclin-dependent kinase inhibitor</td>
</tr>
<tr>
<td>BCL2</td>
<td>B-cell CLL/lymphoma 2</td>
</tr>
<tr>
<td>BAX</td>
<td>BCL2-associated X protein</td>
</tr>
<tr>
<td>PCNA</td>
<td>proliferating cell nuclear antigen</td>
</tr>
<tr>
<td>MDM2</td>
<td>Mdm2, transformed 3T3 cell...</td>
</tr>
<tr>
<td>APAF1</td>
<td>apoptotic peptidase activating factor</td>
</tr>
<tr>
<td>GADD45A</td>
<td>growth arrest and DNA-damaging...</td>
</tr>
</tbody>
</table>
References

GSEA:
http://www.broadinstitute.org/gsea/

Genepattern:
http://www.broadinstitute.org/cancer/software/genepattern/
ARACNE demo

- ARACNE theory recap
- Java GUI
- Command line tool

Xi Zhao, PhD
Cancer Center for Systems Biology, Stanford University
xi.zhao@stanford.edu
Recap: interactome

- Genes do not function alone
- Construct gene interaction network from genomic data

**e.g Expression matrix**  **Pair-wise correlation matrix**

**Post processing:** e.g. correlation significance

**ARACNE: Mutual Information**

**ARACNE: Data Processing Inequality**
Recap: ARACNE

• Mutual Information (MI):
  • detect non-linear dependencies between a pair of variables \((X, Y)\)

• Data Processing Inequality (DPI):
  • remove weakest link (tends to be indirect interaction) in a triple

\[
\begin{align*}
I_1 &> I_2 > I_3 \\
G2 & \rightarrow G1 \rightarrow G3 \\
& \leftarrow G3 \rightarrow G2
\end{align*}
\]
ARACNE limitations

1. MI has low statistical power although detect non-linear dependencies

   MI “has lower power than distant correlation, ... (it) is sometimes less powerful than Pearson correlation as well, the linear case being particularly worrisome.” (Simon and Tibshirani, 2011 comment on Reshef et al 2011)

2. Can easily turn into Hair balls: Difficulty in interpretation
ARACNE java GUI: Setup

• **Download** ([link](http://wiki.c2b2.columbia.edu/califanolab/index.php/Software/ARACNE))
  - aracne.zip ([link](http://wiki.c2b2.columbia.edu/califanolab/download/ARACNE/aracne.zip))
  - Java JDK 1.5 or newer (compiling issues with 1.7 on Mac?)

• **Set $PATH variable** (on Mac)
  - edit `.bashrc` file (e.g. `vi ~/.bashrc`)
    ```
    export PATH="$(/usr/libexec/java_home)/bin:$PATH"
    ```
  - edit `launch_aracne.sh`
    ```
    java -cp lib/ant.jar:lib/ant-launcher.jar org.apache.tools.ant.launch.Launcher run
    ```

• **Launch from terminal:**
  - `cd <aracne folder>`
  - `$sh launch_aracne.sh`
ARACNE file requirement

- Input file - expression matrix (.exp):
  - TAB-delimited text files

- Output file - adjacency matrix file (.adj):
  - TAB-delimited text files
ARACNE java GUI

- Load “BCell_matrix.exp”
- Construct network around hubs (optional)
- TF list (optional)
- Output: adjacency file “.adj”
- Name the output file (optional)
- Load existing adjacency file “.adj” (optional)
- Filter by p value (optional)
- Filter by DPI tolerance (optional): 0~0.2
- Output: adjacency file “.adj”

<table>
<thead>
<tr>
<th>AffyID</th>
<th>Annotation</th>
<th>BL74</th>
<th>KEMtype-A</th>
<th>Mutu-I-A</th>
<th>Namalwa</th>
<th>ODHtype-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-hum_a</td>
<td>ADRBK2</td>
<td>7123.1</td>
<td>4076.7</td>
<td>3228.3</td>
<td>10169.3</td>
<td>2276.7</td>
</tr>
<tr>
<td>AFFX-HUMIS</td>
<td>STAT1</td>
<td>117.7</td>
<td>19.3</td>
<td>16.7</td>
<td>46.7</td>
<td>26.8</td>
</tr>
<tr>
<td>AFFX-HUMG</td>
<td>GAPD</td>
<td>9279.1</td>
<td>5355.3</td>
<td>2719.8</td>
<td>10551.6</td>
<td>4530.1</td>
</tr>
<tr>
<td>AFFX-HUMG</td>
<td>GAPD</td>
<td>7245.9</td>
<td>4978.3</td>
<td>2368.5</td>
<td>9836.4</td>
<td>3874.3</td>
</tr>
<tr>
<td>AFFX-HUMG</td>
<td>GAPD</td>
<td>6796.9</td>
<td>5931.1</td>
<td>2248.3</td>
<td>9916.5</td>
<td>4150.9</td>
</tr>
<tr>
<td>AFFX-HSACO</td>
<td>ACTB</td>
<td>8042.7</td>
<td>4482.1</td>
<td>3243.1</td>
<td>7457.1</td>
<td>4173.1</td>
</tr>
</tbody>
</table>

Program Output

1827_s_at
1973_s_at
37724_at

TF list (optional)
ARACNE java GUI

- Network visualization
  - Network Browser -> Load -> Draw -> Cytoscape
ARACNE Command Line

Download ARACNE.src.tar.gz

ARACNE options: http://wiki.c2b2.columbia.edu/califanolab/images/9/92/Usage.txt

- i <file> Input gene expression profile dataset
- o <file> Output file name (optional)
- j <file> Existing adjacency matrix (.adj) file
- e <tolerance> DPI tolerance
- p <p-value> P-value for MI threshold (e.g. 1e-7)

$ cd <ARACNE.src/ARACNE folder>

Construct network using all probes provided in BCell_matrix.exp

$.aracne2 -i ../../aracne/Data/BCell_matrix.exp

Construct network around probes provided in myc_probes.txt

$.aracne2 -i ../../aracne/Data/BCell_matrix.exp -s ../../aracne/Data/myc_probes.txt

Load existing adjacency file .adj with DPI & P-value filtering

$.aracne2 -i ../../aracne/Data/BCell_matrix.exp -j ../../aracne/Data/BCell_matrix_k0.139.adj -p 0.001 -e 0.15
References


