Plan

- Basic gene expression profile analysis
- Some issues in gene expression analysis
- Improving reproducibility
- More advanced analysis
  - Small-sample-size analysis
  - Novel principle for childhood ALL relapse prediction
BASIC GENE EXPRESSION PROFILE ANALYSIS
Affymetrix GeneChip

Source: Affymetrix
quartz is washed to ensure uniform hydroxylation across its surface and to attach linker molecules

exposed linkers become deprotected and are available for nucleotide coupling

Source: Affymetrix
Gene expression measurement by Affymetrix GeneChip

Source: Affymetrix
Diagnosis using microarray

Image credit: Affymetrix
Application: Disease subtype diagnosis

[Diagram showing a grid with samples and genes, where some samples are labeled as benign and others as malignant.]
Application: Drug action detection

Which group of genes are the drug affecting on?
Typical analysis workflow

- Gene expression data collection
- DE gene selection by, e.g., t-statistic
- Classifier training based on selected DE genes
- Apply the classifier for diagnosis of future cases

Terminology: DE gene = differentially expressed gene

Signal Selection Basic Idea
- Choose a signal with low intra-class distance
- Choose a signal with high inter-class distance

Image credit: Golub et al., Science, 286:531–537, 1999

If you don’t remember this, you should go back to revise your CS2220 notes 😊
You can build a gene expression profile classifier in a simple or in a more complex way.

- **Parallel-multiclass classification scheme**

Next, we take an example to demonstrate the scores used by PCL. A BCR-ABL test sample contained almost all of the top-20 BCR-ABL discriminators. So, a score of 19.6 was assigned to it. Several top-20 OTHERS discriminators together with some beyond the top-20 list were also contained in this test sample. So, another score of 6.97 was assigned. This test sample did not contain any discriminators of E2A-PBX1, Hyperdip>50, or T-ALL. So, the scores are as follows:

<table>
<thead>
<tr>
<th>subtype</th>
<th>BCR-ABL</th>
<th>E2A-PBX1</th>
<th>Hyperdip&gt;50</th>
<th>T-ALL</th>
<th>MLL</th>
<th>TEL-AML1</th>
<th>OTHERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>19.63</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.71</td>
<td>2.96</td>
<td>6.97</td>
</tr>
</tbody>
</table>


- **Tree-structured classification scheme**

Childhood ALL Subtype Diagnosis Workflow

A tree-structured diagnostic workflow was recommended by our doctor collaborator.
Hierarchical clustering

PCA plots

SOME ISSUES IN GENE EXPRESSION ANALYSIS
Some headaches

• Natural fluctuations of gene expression in a person

• Noise in experimental protocols
  – Numbers mean different things in different batches
  – Numbers mean different things in data obtained from different platforms

⇒ Selected genes may not be meaningful
  – Different genes get selected in different experiments
Natural fluctuations & expt noise

Intrinsic & extrinsic noise

Measurement errors


Sometimes, a gene expression study may involve batches of data collected over a long period of time...

### Batch effects

- Samples from different batches are grouped together, regardless of subtypes and treatment response

Image credit: Difeng Dong’s PhD dissertation, 2011
Percentage of overlapping genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

<table>
<thead>
<tr>
<th>Datasets</th>
<th>DEG</th>
<th>POG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prostate Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td><strong>Lung Cancer</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td><strong>DMD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>Top 50</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Top100</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

“Most random gene expression signatures are significantly associated with breast cancer outcome”

BETTER REPRODUCIBILITY OF GENE EXPRESSION PROFILE ANALYSIS
Individual genes

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples

• How many genes on a microarray are expected to perfectly correlate to these samples?

• Prob(a gene is correlated) = 1/2^6
• # of genes on array = 25,000
⇒ E(# of correlated genes) = 390

⇒ Many false positives
• These cannot be eliminated based on pure statistics!
Group of genes

• Suppose
  – Each gene has 50% chance to be high
  – You have 3 disease and 3 normal samples
• What is the chance of a group of 5 genes being perfectly correlated to these samples?

• Prob(group of genes correlated) = \( \frac{1}{2^6} \)^5
  – Good, << \( \frac{1}{2^6} \)
• # of groups = \( \binom{25000}{5} \)
  ⇒ \( \mathbb{E}(\text{# of groups of genes correlated}) = \binom{25000}{5} \star (\frac{1}{2^6})^5 = 7.58 \times 10^{10} \)

⇒ Even more false positives?
• Perhaps no need to consider every group
Gene-regulatory circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype
Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

- Prob(group of genes correlated) = \((1/2^6)^5\)
  - Good, << 1/2^6
- \# of groups = \(\binom{25000}{5}\)
- \(\implies\) E(# of groups of genes correlated) = \(\binom{25000}{5} \cdot (1/2^6)^5 = 7.50 \times 10^{10}\)

\(\implies\) Even more false positives?
- Perhaps no need to consider every group
Towards more meaningful genes

• ORA
  – Khatri et al
  – *Genomics*, 2002

• FCS
  – Pavlidis & Noble
  – PSB 2002

• GSEA
  – Subramanian et al
  – *PNAS*, 2005

• PFSNet
  – Lim & Wong
  – *Bioinformatics*, 2014

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis
Overlap analysis: ORA

ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (e.g., genes whose t-statistic meets the 5% significance threshold of t-test), and checking the significance of the size of the intersection using the hypergeometric test.

Disappointing performance

upregulated in DMD

![Graph showing subnetwork agreement vs sample size (N)]

- DMD gene expression data
  - Pescatori et al., 2007
  - Haslett et al., 2002
- Pathway data
  - PathwayAPI, Soh et al., 2010
Time for Exercise #1

- Discuss why ORA performs so poorly in selecting differentially expressed genes
**Direct-group analysis: FCS**

\[ \frac{1}{n} \sum_{k=1}^{n} -\log(p_k) \]

- **GO Class 1** → **Score 1** → **Significant Class 1**
- **GO Class 2** → **Score 2** → **Non Significant Class 2**
- **GO Class N** → **Score 3** → **Significant Class N**

- Score of a pathway C = average of log-transformed p-values of genes in C.
- Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C.

Where will FCS be in comparison to ORA below? Why?

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
Direct-group analysis: GSEA

Where will GSEA be in comparison to ORA and FCS? Why?

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
Time for Exercise #2

• What is the key difference between the null hypothesis underlying FCS and GSEA?
PFSNet: Exploiting subnetworks

- Induce subnetworks from pathways by considering only genes highly expressed in majority of patients in any class

\[
\beta_1^*(g_i) = \sum_{p_j \in D} \frac{f_s(e_{g_i,p_j})}{|D|} \\
\beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{f_s(e_{g_i,p_j})}{|\neg D|}
\]

- For an irrelevant subnetwork \( S \), the two scores above for each patient \( P_k \) should be roughly equal, regardless of his class

- Do a paired t-test to decide whether \( S \) is relevant
  - Get null distribution by permuting class labels
Where will PFSNet be in comparison to ORA, FCS, and GSEA? Why?

DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010
What have we learned?

• Common headaches in gene expression analysis
  – Natural fluctuation, protocol noise, batch effect

• Use of biological background info to tame false positives

• Overlap analysis → direct-group analysis → network-based analysis

• Subnetwork-based methods yield more consistent and larger disease subnetworks
From pathways to models, From static to dynamic: A couple of quite recent papers that are worth your leisure reading…


Must read


Good to read


This part of the lecture is show you an example of more advanced forms of gene expression analysis

GENE EXPRESSION PROFILE ANALYSIS WHEN SAMPLE SIZE IS EXTREMELY SMALL
Recall the 3 issues of ORA

Issue #2 with ORA

- It relies on a pre-determined list of DE genes
- This list is sensitive to the test statistic used and to the significance threshold used
- This list is unstable regardless of the threshold used when sample size is small

Issue #3 with ORA

- It tests whether the entire pathway is significantly differentially expressed
- If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch
Outline

• A new idea for a valid null hypothesis: ORA-Paired

• Paired test on subnetworks: NEA-Paired

• Achieving a quantum leap: ESSNet

Main reference for this work

ORA-Paired: Paired test and new null hypothesis

- Let $g_i$ be genes in a given pathway $P$
- Let $p_j$ be a patient
- Let $q_k$ be a normal

Let $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$

Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- **Issue #1 is solved**
  - Null hypothesis is “Pathway $P$ is irrelevant to the difference between patients and normals, and the genes in $P$ behave similarly in patients and normals”

- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes

- **Issue #3 is unsolved**

- **Is sample size now larger?**
  - $|\text{patients}| \times |\text{normals}| \times |\text{genes in } P|$
Testing the null hypothesis

“Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals”

• Method #1
  – T-test w/ a conservative degree of freedom
    • E.g., # normals + # patients

• Method #2
  – By the null hypothesis, a dataset and any of its class-label permutations are exchangeable
    ⇒ Get null distribution by class-label permutations
    • Only for large-size sample

• Method #3
  – Modified null hypothesis
    • “Pathway P induces gene-gene correlations, and genes in P behave according to these gene-gene correlations;
    • P is irrelevant to the diff betw patients and normals and so, genes in P behave similarly in patients and normals”
    ⇒ Get null distribution using datasets that conserve gene-gene correlations in the original dataset
    • E.g., array rotation
Similar to PFSNet, good but not great.
NEA-Paired:
Paired test on subnetworks

- Given a pathway $P$
- Let each node and its immediate neighbourhood in $P$ be a subnetwork
- Apply ORA-Paired on each subnetwork individually

- Issues #1 & #2 are solved as per ORA-Paired

- Issue #3 is partly solved
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in a fragmented way
Much better performance
ESSNet: Larger subnetworks

- Compute the average rank of a gene based on its expression level in patients in any class
- Use the top $\alpha\%$ to extract large connected components in pathways
- Test each component using ORA-Paired

- Gene rank is very stable
- Issues #1 - #3 solved
Fantastic performance
More datasets tested
ESSNet is unlikely to report junk

Table 4.2: Average number of subnetworks predicted by ESSNet over the sample sizes (N); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

<table>
<thead>
<tr>
<th>Sample size (N)</th>
<th>DMD</th>
<th>ALL</th>
<th>BCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8.2/13.4</td>
<td>7.0/11.9</td>
<td>4.8/12.6</td>
</tr>
<tr>
<td>3</td>
<td>11.1/15.9</td>
<td>11.3/17.9</td>
<td>5.0/11.7</td>
</tr>
<tr>
<td>4</td>
<td>13.18/16.5</td>
<td>11.9/15.9</td>
<td>6.2/10.4</td>
</tr>
<tr>
<td>5</td>
<td>14.2/16.7</td>
<td>14.6/18.3</td>
<td>7.9/12.7</td>
</tr>
<tr>
<td>6</td>
<td>15.14/17.6</td>
<td>14.9/18.0</td>
<td>11.0/15.7</td>
</tr>
<tr>
<td>7</td>
<td>15.2/17.4</td>
<td>16.1/19.2</td>
<td>12.9/17.5</td>
</tr>
<tr>
<td>8</td>
<td>15.4/17.5</td>
<td>16.2/19.0</td>
<td>15.3/20.4</td>
</tr>
<tr>
<td>9</td>
<td>16.6/18.8</td>
<td>17.0/19.8</td>
<td>15.8/20.8</td>
</tr>
<tr>
<td>10</td>
<td>17.6/19.7</td>
<td>17.3/19.7</td>
<td>16.2/20.8</td>
</tr>
</tbody>
</table>
Time for Exercise #3

- Discuss how you would design an experiment to test whether a gene-selection method controls false positives well
ESSNet also dominates when sample size is large

**Table 4.3**: Number of subnetworks predicted by the various methods on a full dataset where the null distribution is computed using array rotation (rot), class-label swapping (cperm) and gene swapping (gswap); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

<table>
<thead>
<tr>
<th>Method</th>
<th>DMD</th>
<th>ALL</th>
<th>BCR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rot cperm</td>
<td>rot cperm</td>
<td>rot cperm</td>
</tr>
<tr>
<td>ESSNet</td>
<td>20/23 13/15</td>
<td>22/24 25/27</td>
<td>24/29 30/32</td>
</tr>
<tr>
<td>NEA-paired</td>
<td>77/98 91/115</td>
<td>140/163 109/119</td>
<td>176/192 37/43</td>
</tr>
<tr>
<td>ORA-paired</td>
<td>30/62 30/62</td>
<td>34/74 34/74</td>
<td>53/99 53/99</td>
</tr>
<tr>
<td>ORA-hypergeo</td>
<td>20/46 41/141</td>
<td>24/60 48/73</td>
<td>4/14 32/166</td>
</tr>
<tr>
<td>GSEA</td>
<td>cperm gswap</td>
<td>cperm gswap</td>
<td>cperm gswap</td>
</tr>
<tr>
<td></td>
<td>23/64 24/69</td>
<td>8/52 17/48</td>
<td>7/57 5/46</td>
</tr>
</tbody>
</table>

Do ESSNet results agree on small datasets vs big datasets?

- Use ESSNet’s results on entire datasets as the benchmark to evaluate ESSNet’s results on small subsets of the datasets.
- The precision (i.e., agreement) is superb, though some subnetworks are missed when smaller datasets are analysed.
This part of the lecture is show you another example of more advanced forms of gene expression analysis

A NOVEL PRINCIPLE FOR CHILDHOOD ALL RELAPSE PREDICTION
Childhood acute lymphoblastic leukemia

- The most common cancer in children
  - 3,000 new cases in US
  - 2,000 new cases in ASEAN countries

- 80% achieve long-term relapse-free survival, but
  - 20% relapse and eventually die
  - Large fraction of them suffer severe side effects

⇒ Predict relapse early and treat more aggressively
Outline

• Background
• Hypotheses
• Framework
• Methodologies
  – Data Preparation
  – Model Construction
  – Relapse Prediction
• Validation
• Conclusion

Main reference for this work
• Difeng Dong, "Relapse Prediction in Childhood Acute Lymphoblastic Leukemia by Time-Series Gene Expression Profiling", PhD thesis, November 2011, National University of Singapore
Contemporary ALL treatment framework

- Treatment A: Intensive Treatment
- Treatment B: Moderate Treatment
Previous work

• Correlate GEP to childhood ALL subtypes
  – Identified subtype-based genetic signatures
  – Diagnostic accuracy is >95%, better than routine diagnostic methods

Yeoh et al. Cancer Cell, 2002

What does intensity of genetic signature means?
Hypotheses

• Treatment gradually removes leukemic cells in patient

• Diagnostic GEP captures leukemic subtype signature

• Hypothesis 1: Time-series GEP captures reduction of leukemic cells during treatment

• Hypothesis 2: Poor genetic response suggests high risk of relapse
Framework

• Time-series GEP data preparation (normalization)

• H1: Time-series GEP captures reduction of leukemic cells during treatment
  – Unsupervised hierarchical clustering
  – Signature dissolution analysis
  – Genetic status shifting (GSS) model

• H2: Poor genetic response ➔ high risk of relapse
  – Prediction based on GSS distance

• Validation in independent datasets
GEP data preparation

- 96 patients, 10 relapses vs 86 remissions
- GEP collected on 4 time points, D0, D8, D15, D33, a matrix of >30,000 genes * >300 samples
- Data generated by MAS5.0
GEP data normalization

- Scaling factor $>20 \rightarrow$ Remove
- Noise mainly in low-expression genes
- Genes with $> 70\%$ absent calls $\rightarrow$ Remove
- 4,736 genes remain
- Perform quantile normalization
Framework

• **Time-series GEP data preparation (normalization)**

• **H1:** Time-series GEP captures reduction of leukemic cells during treatment
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model

• **H2:** Poor genetic response $\rightarrow$ high risk of relapse
  - Prediction based on GSS distance

• Validation in independent datasets
Unsupervised clustering

- Top 10% of genes with largest variance across whole dataset
  - 1,474 genes
  - Noise mainly in low-expression genes

- Unsupervised hierarchical clustering on patients
  - Pearson's correlation
  - Completed linkage
Unsupervised Clustering: Results
Signature dissolution analysis

- How do intensity of genetic signatures change during treatment?

Yeoh et al. *Cancer Cell*, 2002
Signature dissolution analysis

- Consider the 3 largest subtypes
  - TEL-AML1, n = 26
  - T-ALL, n = 12
  - Hyperdiploid>50, n = 12

- Select genetic signature genes for them
  - Organize diagnostic samples into Subtype A vs ~A
  - Only consider genes highly expressed in A
  - Pick 20 most diff expressed genes by t-test
    - The selected signature genes was validated by testing their prediction accuracy on public data
    - Accuracy achieved: 95%
Signature dissolution analysis

- Red is highly expressed
- Green is lowly expressed
Global genetic status shifting (GSS) model

- **Select drug responsive genes**
  - Diff expressed genes betw D0 & D8 by t-test \((q<0.0001)\)
  - >2 fold change betw D0 and D8
  \(\Rightarrow\) 461 up- and 99 down-regulated genes

- **Apply principal component analysis**
  - Genes are considered as features
  - PCA
  - Each point is a sample
Global Genetic Status Shifting Model

(a)

(b)

<table>
<thead>
<tr>
<th>PC</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>49.08%</td>
<td>7.56%</td>
<td>5.19%</td>
<td>3.64%</td>
<td>2.10%</td>
<td>1.81%</td>
<td>1.61%</td>
<td>1.35%</td>
<td>72.34%</td>
</tr>
</tbody>
</table>
Verifying the global GSS model
GO and Pathway Ingenuity Analysis on drug-responsive genes in the global GSS model

- **UP:** Reconstruction of immune system and restoration of normal hematogenesis
- **DOWN:** Cell development and DNA synthesis
- **DOWN:** Negative regulation of apoptosis
Framework

- Time-series GEP data preparation (normalization)

- H1: Time-series GEP captures reduction of leukemic cells during treatment
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model

- H2: Poor genetic response \(\Rightarrow\) high risk of relapse
  - Prediction based on GSS distance

- Validation in independent datasets
Genetic status shifting distance

DC: Disease Centroid
NC: NBM Centroid

ASD = s3  (Absolute shifting distance)
ESD = s2  (Effective shifting distance)
ESR = s2/s1  (Effective shifting ratio)
## Relapse prediction by ESD

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56_KL464</td>
<td>-2.21</td>
<td>30</td>
<td>29_KL439</td>
<td>26.20</td>
<td>59</td>
<td>05_KL354</td>
<td>52.56</td>
</tr>
<tr>
<td>2</td>
<td>77_KL401</td>
<td>-1.61</td>
<td>31</td>
<td>49_R432</td>
<td>26.77</td>
<td>60</td>
<td>80_KL423</td>
<td>53.33</td>
</tr>
<tr>
<td>3</td>
<td>33_R247</td>
<td>-0.46</td>
<td>32</td>
<td>83_KL457</td>
<td>29.13</td>
<td>61</td>
<td>75_KL385</td>
<td>54.31</td>
</tr>
<tr>
<td>4</td>
<td>19_KL205</td>
<td>-0.21</td>
<td>33</td>
<td>04_KL322</td>
<td>30.64</td>
<td>62</td>
<td>69_KL313</td>
<td>54.40</td>
</tr>
<tr>
<td>5</td>
<td>45_R194</td>
<td>2.13</td>
<td>34</td>
<td>74_KL383</td>
<td>31.52</td>
<td>63</td>
<td>64_KKH29</td>
<td>55.12</td>
</tr>
<tr>
<td>6</td>
<td>59_R281</td>
<td>2.50</td>
<td>35</td>
<td>60_KKH30</td>
<td>32.88</td>
<td>64</td>
<td>93_R337</td>
<td>55.35</td>
</tr>
<tr>
<td>7</td>
<td>97_R208</td>
<td>2.94</td>
<td>36</td>
<td>38_KL218</td>
<td>33.00</td>
<td>65</td>
<td>26_KL369</td>
<td>56.83</td>
</tr>
<tr>
<td>8</td>
<td>67_KL287</td>
<td>4.63</td>
<td>37</td>
<td>10_R257</td>
<td>33.39</td>
<td>66</td>
<td>14_KKH19</td>
<td>57.79</td>
</tr>
<tr>
<td>9</td>
<td>96_R202</td>
<td>5.14</td>
<td>38</td>
<td>40_KL430</td>
<td>33.55</td>
<td>67</td>
<td>28_KL375</td>
<td>58.21</td>
</tr>
<tr>
<td>10</td>
<td>11_R280</td>
<td>5.90</td>
<td>39</td>
<td>86_KL509</td>
<td>34.48</td>
<td>68</td>
<td>62_KKH22</td>
<td>60.55</td>
</tr>
<tr>
<td>11</td>
<td>27_KL374</td>
<td>6.31</td>
<td>40</td>
<td>51_KL461</td>
<td>34.83</td>
<td>69</td>
<td>21_KL300</td>
<td>60.80</td>
</tr>
<tr>
<td>12</td>
<td>70_KL320</td>
<td>6.64</td>
<td>41</td>
<td>79_KL421</td>
<td>35.43</td>
<td>70</td>
<td>01_KKH18</td>
<td>62.07</td>
</tr>
<tr>
<td>13</td>
<td>41_KL441</td>
<td>7.86</td>
<td>42</td>
<td>94_R354</td>
<td>35.66</td>
<td>71</td>
<td>32_R233</td>
<td>67.35</td>
</tr>
<tr>
<td>14</td>
<td>20_KL274</td>
<td>8.91</td>
<td>43</td>
<td>24_KL328</td>
<td>35.78</td>
<td>72</td>
<td>34_R256</td>
<td>67.95</td>
</tr>
<tr>
<td>15</td>
<td>82_KL454</td>
<td>9.53</td>
<td>44</td>
<td>47_R334</td>
<td>37.35</td>
<td>73</td>
<td>68_KL304</td>
<td>70.06</td>
</tr>
<tr>
<td>16</td>
<td>13_R410</td>
<td>11.45</td>
<td>45</td>
<td>30_KL444</td>
<td>38.26</td>
<td>74</td>
<td>18_KKH28</td>
<td>70.41</td>
</tr>
<tr>
<td>17</td>
<td>39_KL395</td>
<td>11.49</td>
<td>46</td>
<td>57_KL535</td>
<td>39.40</td>
<td>75</td>
<td>36_R343</td>
<td>71.95</td>
</tr>
<tr>
<td>18</td>
<td>17_KKH27</td>
<td>12.64</td>
<td>47</td>
<td>73_KL381</td>
<td>39.51</td>
<td>76</td>
<td>58_KL543</td>
<td>71.96</td>
</tr>
<tr>
<td>19</td>
<td>35_R313</td>
<td>13.76</td>
<td>48</td>
<td>66_KL247</td>
<td>39.69</td>
<td>77</td>
<td>08_KL456</td>
<td>72.39</td>
</tr>
<tr>
<td>20</td>
<td>55_KL419</td>
<td>18.60</td>
<td>49</td>
<td>16_KKH21</td>
<td>40.05</td>
<td>78</td>
<td>12_R297</td>
<td>74.21</td>
</tr>
<tr>
<td>21</td>
<td>99_KL416</td>
<td>19.83</td>
<td>50</td>
<td>84_KL458</td>
<td>40.12</td>
<td>79</td>
<td>88_KL544</td>
<td>75.25</td>
</tr>
<tr>
<td>22</td>
<td>23_KL321</td>
<td>20.37</td>
<td>51</td>
<td>37_R355</td>
<td>41.68</td>
<td>80</td>
<td>72_KL377</td>
<td>76.58</td>
</tr>
<tr>
<td>23</td>
<td>50_KL360</td>
<td>20.80</td>
<td>52</td>
<td>25_KL357</td>
<td>42.14</td>
<td>81</td>
<td>44_KL541</td>
<td>79.45</td>
</tr>
<tr>
<td>24</td>
<td>90_R253</td>
<td>21.40</td>
<td>53</td>
<td>15_KKH20</td>
<td>43.35</td>
<td>82</td>
<td>87_KL522</td>
<td>79.91</td>
</tr>
<tr>
<td>25</td>
<td>61_KKH13</td>
<td>21.67</td>
<td>54</td>
<td>52_R252</td>
<td>45.31</td>
<td>83</td>
<td>89_R245</td>
<td>85.37</td>
</tr>
<tr>
<td>26</td>
<td>43_KL536</td>
<td>23.00</td>
<td>55</td>
<td>85_KL485</td>
<td>45.82</td>
<td>84</td>
<td>63_KKH25</td>
<td>85.67</td>
</tr>
<tr>
<td>27</td>
<td>65_KL224</td>
<td>23.30</td>
<td>56</td>
<td>92_R332</td>
<td>49.15</td>
<td>85</td>
<td>78_KL412</td>
<td>92.38</td>
</tr>
<tr>
<td>28</td>
<td>07_KL417</td>
<td>24.00</td>
<td>57</td>
<td>98_KL387</td>
<td>49.17</td>
<td>86</td>
<td>95_R431</td>
<td>104.57</td>
</tr>
<tr>
<td>29</td>
<td>42_KL507</td>
<td>24.00</td>
<td>58</td>
<td>48_R339</td>
<td>49.26</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Comparison with other clinical & computational protocols

<table>
<thead>
<tr>
<th>Method</th>
<th>Prognostic Feature</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holleman-DT</td>
<td>D0 GEP</td>
<td>60.00%</td>
<td>69.51%</td>
<td>64.76%</td>
</tr>
<tr>
<td>Holleman-NB</td>
<td>D0 GEP</td>
<td>60.00%</td>
<td>69.51%</td>
<td>64.76%</td>
</tr>
<tr>
<td>Holleman-SVM</td>
<td>D0 GEP</td>
<td>60.00%</td>
<td>69.51%</td>
<td>64.76%</td>
</tr>
<tr>
<td>Bhojwani</td>
<td>D0 GEP</td>
<td>20.00%</td>
<td>79.27%</td>
<td>49.63%</td>
</tr>
<tr>
<td>NCI</td>
<td>D0 GEP</td>
<td>80.00%</td>
<td>58.14%</td>
<td>69.07%</td>
</tr>
<tr>
<td>Cytogenetics</td>
<td>Diagnostic Cytogenetics</td>
<td>30.00%</td>
<td>94.19%</td>
<td>62.09%</td>
</tr>
<tr>
<td>MRD-D33</td>
<td>D33 MRD</td>
<td>77.78%</td>
<td>54.12%</td>
<td>65.95%</td>
</tr>
<tr>
<td>D8 Response</td>
<td>D8 Blast Count</td>
<td>30.00%</td>
<td>85.53%</td>
<td>57.76%</td>
</tr>
<tr>
<td>ASD-D8</td>
<td>D0 and D8 GEP</td>
<td>90.00%</td>
<td>73.68%</td>
<td>81.84%</td>
</tr>
<tr>
<td>ESD-D8</td>
<td>D0 and D8 GEP</td>
<td>100.00%</td>
<td>75.00%</td>
<td>87.50%</td>
</tr>
<tr>
<td>ESR-D8</td>
<td>D0 and D8 GEP</td>
<td>90.00%</td>
<td>73.68%</td>
<td>81.84%</td>
</tr>
</tbody>
</table>
Framework

- Time-series GEP data preparation (normalization)

- **H1**: Time-series GEP captures reduction of leukemic cells during treatment
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model

- **H2**: Poor genetic response ➔ high risk of relapse
  - Prediction based on GSS distance

- Validation in independent datasets
Validation on an independent dataset

- 8 childhood ALL patients from Europe
- GEP on D0, D8, and D15
- Standard data preprocessing
- Use the same drug-responsive genes as the global GSS model
- Apply PCA and use ESD to make prediction
Recall... so threshold is ESD = 24

<table>
<thead>
<tr>
<th>RANK</th>
<th>SAMPLE</th>
<th>ESD-D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56_KL464</td>
<td>-2.21</td>
</tr>
<tr>
<td>2</td>
<td>77_KL401</td>
<td>-1.51</td>
</tr>
<tr>
<td>3</td>
<td>33_R247</td>
<td>-0.46</td>
</tr>
<tr>
<td>4</td>
<td>19_KL205</td>
<td>-0.21</td>
</tr>
<tr>
<td>5</td>
<td>45_R194</td>
<td>2.13</td>
</tr>
<tr>
<td>6</td>
<td>59_R281</td>
<td>2.50</td>
</tr>
<tr>
<td>7</td>
<td>97_R208</td>
<td>2.94</td>
</tr>
<tr>
<td>8</td>
<td>67_KL287</td>
<td>4.63</td>
</tr>
<tr>
<td>9</td>
<td>96_R202</td>
<td>5.14</td>
</tr>
<tr>
<td>10</td>
<td>11_R280</td>
<td>5.90</td>
</tr>
<tr>
<td>11</td>
<td>27_KL374</td>
<td>6.31</td>
</tr>
<tr>
<td>12</td>
<td>70_KL320</td>
<td>6.64</td>
</tr>
<tr>
<td>13</td>
<td>41_KL441</td>
<td>7.86</td>
</tr>
<tr>
<td>14</td>
<td>20_KL274</td>
<td>8.91</td>
</tr>
<tr>
<td>15</td>
<td>82_KL454</td>
<td>9.53</td>
</tr>
<tr>
<td>16</td>
<td>13_R410</td>
<td>11.45</td>
</tr>
<tr>
<td>17</td>
<td>39_KL395</td>
<td>11.49</td>
</tr>
<tr>
<td>18</td>
<td>17_KKH27</td>
<td>12.64</td>
</tr>
<tr>
<td>19</td>
<td>35_R313</td>
<td>13.76</td>
</tr>
<tr>
<td>20</td>
<td>55_KL419</td>
<td>18.60</td>
</tr>
<tr>
<td>21</td>
<td>99_KL416</td>
<td>19.83</td>
</tr>
<tr>
<td>22</td>
<td>23_KL321</td>
<td>20.37</td>
</tr>
<tr>
<td>23</td>
<td>50_KL360</td>
<td>20.80</td>
</tr>
<tr>
<td>24</td>
<td>90_R253</td>
<td>21.40</td>
</tr>
<tr>
<td>25</td>
<td>61_KKH13</td>
<td>21.67</td>
</tr>
<tr>
<td>26</td>
<td>43_KL536</td>
<td>23.00</td>
</tr>
<tr>
<td>27</td>
<td>65_KL224</td>
<td>23.30</td>
</tr>
<tr>
<td>28</td>
<td>07_KL417</td>
<td>24.00</td>
</tr>
<tr>
<td>29</td>
<td>42_KL507</td>
<td>24.00</td>
</tr>
</tbody>
</table>
### Result on the independent dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Subtype</th>
<th>ESD</th>
<th>Chance of adverse event</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>861-M</td>
<td>ALL</td>
<td>BCR-ABL</td>
<td>-4.67</td>
<td>Very high risk</td>
<td>100%</td>
</tr>
<tr>
<td>1021-M</td>
<td>ALL</td>
<td>BCR-ABL</td>
<td>1.86</td>
<td>Very high risk</td>
<td>90%</td>
</tr>
<tr>
<td>961-M</td>
<td>ALL</td>
<td>HYPERDIP</td>
<td>6.26</td>
<td>High risk</td>
<td>60%</td>
</tr>
<tr>
<td>906-M</td>
<td>ALL</td>
<td>OTHERS</td>
<td>17.16</td>
<td>High risk</td>
<td>30%</td>
</tr>
<tr>
<td>980-M</td>
<td>ALL</td>
<td>OTHERS</td>
<td>30.84</td>
<td>Intermediate to low risk</td>
<td>0%</td>
</tr>
<tr>
<td>824-M</td>
<td>ALL</td>
<td>TEL-AML1</td>
<td>39.16</td>
<td>Intermediate to low risk</td>
<td>0%</td>
</tr>
<tr>
<td>1003-M</td>
<td>ALL</td>
<td>HYPERDIP</td>
<td>41.48</td>
<td>Intermediate to low risk</td>
<td>0%</td>
</tr>
<tr>
<td>940-M</td>
<td>ALL</td>
<td>TEL-AML1</td>
<td>59.07</td>
<td>Intermediate to low risk</td>
<td>0%</td>
</tr>
</tbody>
</table>
Validation on AML Datasets: How general is the principle of the GSS Model

- 8 acute myeloid leukemia (AML) patients
- Similar treatment philosophy but much lower long-term event-free survival rate (40%)
- Unsynchronized GEPs betw D0 and D60
- Standard data preprocessing
- Select drug-responsive genes by MILE-AML vs MILE-NBM
- PCA and use ASD and ESD to make prediction
## Results on AML dataset

| Rank | SAMPLE   | ASD  | Outcome | | Rank | Sample   | ESD   | Outcome |
|------|----------|------|---------||------|----------|-------|---------|
| 1    | R318-D5  | 0.28 | R       | | 1    | R318-D33 | -11.03| R       |
| 2    | KL473-D32| 3.04 | R       | | 2    | R318-D5  | 0.04  | R       |
| 3    | KL343-D36| 4.33 | R       | | 3    | KL473-D32| 2.83  | R       |
| 4    | KL448-D17| 8.11 | R       | | 4    | KL343-D36| 3.34  | R       |
| 5    | KL505-D14| 10.61| R       | | 5    | KL448-D17| 6.99  | R       |
| 6    | R474-D33 | 11.52| R       | | 6    | KL505-D14| 10.33 | R       |
| 7    | R318-D33 | 20.10| R       | | 7    | R474-D33 | 11.31 | R       |
| 8    | R474-D60 | 25.67|         | | 8    | R474-D60 | 25.62 |         |
| 9    | KL336-D31| 27.14|         | | 9    | KL336-D31| 26.65 |         |
| 10   | KL505-D45| 31.07|         | | 10   | KL505-D45| 31.04 |         |
| 11   | KKH14-D36| 35.61|         | | 11   | KKH14-D36| 35.61 |         |
| 12   | KL448-D51| 39.71|         | | 12   | KL448-D51| 39.67 |         |
Fast forward to 2014…

• GSS is now called “Effective Response Metric” (ERM) and has been tested on 8 new batches of 181 patients

• Difference in the shift of GEP towards normal remains true

(b) Shift of time-series GEP towards normal centroid
Effectiveness of GSS/ERM

- Patients with minimal changes in GEP after 8 days of therapy have 4x higher risk of relapse
  - Good ERM \( \rightarrow \) 14.5% relapse
  - Poor ERM \( \rightarrow \) 44.5%

- Better than Day33 MRD
  - Day33 MRD –ve patients
    - Good ERM \( \rightarrow \) 4.4% relapse
    - Poor ERM \( \rightarrow \) 21.7% (5x more)
  - Day33 MRD +ve patients
    - Good ERM \( \rightarrow \) 29.2% relapse
    - Poor ERM \( \rightarrow \) 62.8% (2x more)

- Better than cytogenetic
  - Favourable cytogenetic patients
    - Good ERM \( \rightarrow \) 2.4% relapse
    - Poor ERM \( \rightarrow \) 36.4% (18x more)
ERM vs MRD33

<table>
<thead>
<tr>
<th></th>
<th>ERM</th>
<th>MRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>3/60</td>
<td>81</td>
</tr>
<tr>
<td>Bad</td>
<td>5/9</td>
<td>46</td>
</tr>
</tbody>
</table>

n: n # relapse
m: m # cases

Actual

<table>
<thead>
<tr>
<th></th>
<th>ERM</th>
<th>MRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal</td>
<td>91</td>
<td>81</td>
</tr>
<tr>
<td>Bad</td>
<td>36</td>
<td>46</td>
</tr>
</tbody>
</table>

Acc = 68%
Sens = 68%
Spec = 80%

Acc = 77%
Sens = 72%
Spec = 65%
Acknowledgements

• Much of this lecture is based on the works of my past/current students
  – Koh Chuan Hock (Ah Fu)
  – Donny Soh
  – Dong Difeng
  – Kevin Lim