nCounter Assay

Automated Process

nCounter Prep Station

Hybridize Reporter to RNA

Remove excess reporters

Bind reporter to surface

Immobilize and align reporter

Image surface

Count codes

slides courtesy of Sean Ferree, Nanostring Technologies
nCounter Assay

Automated Process

Hybridize Reporter to RNA ➔ Remove excess reporters ➔ Bind reporter to surface ➔ Immobilize and align reporter ➔ Image surface ➔ Count codes

slides courtesy of Sean Ferree, Nanostring Technologies

One coded reporter = 1 mRNA = 1 gene
24/28 Samples assayed by **qPCR (FFPE)** and **NanoString (FFPE)** and **Microarray (Fresh Frozen)** were grouped together using the PAM50 genes (Bernard, Perou and Ellis, unpublished).
Current Gene Expression Technologies

1. DNA microarrays
   - Moderate precision covering ALL genes, fast, and inexpensive
     - long oligos (60-70mers/Agilent/Nimblegen) – 1-2 oligos/gene
     - short oligos (~25mers/Affymetrix) – 5-10 oligos/gene
     - cDNA arrays (100-1000bp PCR products/Stanford) – 1-2 cDNA clones/gene

2. Quantitative RT-PCR
   - high precision, large dynamic range, fast, and inexpensive
     - smaller number of genes assayed (compatible with Formalin-fixed, Paraffin Embedded Materials (FFPE), which is the medical standard of care)

3. Nanostring nCounter
   - high precision, large dynamic range, fast, and moderate expense
     - single mRNA molecule counting capabilities (compatible with FFPE)

4. Illumina-based RNA-sequencing
   - Gene expression via Massively Parallel Sequencing (MPS)
     - high precision, large dynamic range, not fast, and expensive
     - compatible with FFPE
Illumina Sequencing Technology
Reversible Terminator Chemistry Foundation

DNA/cDNA (0.1-1 ug total RNA)

Sample preparation

Cluster growth

Sequencing

1 2 3 4 5 6 7 8 9

Image acquisition

Base calling

slides courtesy of Gary Schroth, Illumina Inc.
Sequencing with the Illumina Hiseq2500

~300-400 Million Clusters Per Lane (16 lanes / run)
(2x100bp read = ~1,000,000,000,000 bases / run)

slides courtesy of Gary Schroth, Illumina Inc.
1. Spot X = Gene X
2. Determine signal intensity at Spot X
3. Signal intensity value is relative measure of gene expression, which comes right out of the machine
4. Process is repeated for all spots, thus giving the complete data file

1. Spot X = ??? (have sequence 50-100bp)
2. Map all sequences/spots to human genome, or mRNA transcript library, etc.
3. for each gene/transcript, count the number of reads/sequences that “mapped” to it
4. determine “RPKM” value = Reads Per Kilobase Per Million mapped reads, or RSEM value (does isoforms)
5. Process is performed for all reads, giving a value of expression for each transcript
6. Mapping can be done at exon level also
Read Alignment to the Genome

Read

ATGCCATTACACAGCGA

Human Genome Reference

... CGATCTAACGTAGCTAGCTAGCTAGCTAGCTAGCATGCCATTACACAGCGAACAGGGAGCTTAGGCGC...

- GTAGCTAGCTAGCTAGC
- CTAGCTAGCTAGCTAGC
- CGATCTAACGTAGCTAGC
- GAACAGGGAGCTTAGG
- ACAGGGAGCTAGGC
- ATTACACAGCGAACAGG
- GAACAGGGAGCTAAGG
- ACAGGGAGCTAGGCGC

For mRNA this would give us “9 counts” as a measure of expression.
For DNA we have now obtained the sequence of this region and found a SNP
Genome Sequencing Overview

Sequence all fragments, align to reference genome and determine gene sequences = Whole Genome Sequencing (WGS)
Genome Sequencing Overview

Sequence all fragments = Whole Genome Sequencing (WGS)

Sequence all captured fragments, align, and determine gene sequences = Exome Sequencing (~2% of total genomic DNA)
UNC RNA-Seq Bioinformatics Pipeline

- **SRF**
  - Alignment
    - MapSplice
    - STAR
    - BWA
  - FASTQ

- **BAM**
  - COUNTS (exons, transcript level)
    - UNC Tools
    - RSEM data matrix
    - GEO: Gene Expression Omnibus
  - SPLICING (novel transcripts & splice junctions)
    - MapSplice
    - Text
  - MUTATIONS
    - SNVmix, GATK, or SamTools
    - MAF
  - FUSIONS
    - MapSplice
    - Text

- **UNC Transcripts**
  - UNC Transcript database
  - Human Genome database
  - other databases

Joel Parker
Challenges and issues concerning sequencing data

1. these data are genetically identifiable, therefore proper patient consent is a must
2. to put data in, and to take data out, requires proper scientific credentials
3. these genetic data can’t be transferred to a 3rd party, and limitations on use may exist
4. sequencing data sets are huge in size, and thus transporting from one place to another, and long term data storage, are challenges
Benefits of sequencing-based platforms

1. Highly reproducible, great sensitivity, and a large dynamic range

2. no prior knowledge of sequence is needed

3. provides information on alternative splicing

4. provides information on mutations and other types of structural variants (i.e. gene fusions)

5. can identify foreign RNAs or DNAs (viruses or bacteria) depending upon the library protocol
Illumina Hiseq vs. Illumina GA2

Correlation of Log2(RPKM)

The correlation coefficient is 0.994.
Agilent 244K microarray versus mRNA-seq
(Log2 ratio created for mRNA-seq using Common Reference)

Correlation of log2(Ratio) All (13k) genes

Correlation of log2(Ratio) Intrinsic 1628 genes
236 TCGA tumors assayed by Agilent microarrays + mRNA-seq and co-clustered using 13,000 genes

mRNA-seq data was “platform” normalized to Agilent data
Adjustment of systematic microarray data biases

Monica Benito1, Joel Parker2, Quan Du2, Junyuan Wu2, Dong Xiang2, Charles M. Perou3,4,* and J. S. Marron5,*

1 Department of Statistics and Econometrics, University of Carlos III, Madrid, Spain, 2 Lineberger Comprehensive Cancer Center, 3 Department of Genetics and 4 Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599-7264, USA, 5 Department of Molecular Medicine, Karolinska Institutet, S 171 76 Stockholm, Sweden and 6 Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

Received on April 4, 2003; revised on July 3, 2003; accepted on July 12, 2003.

ABSTRACT

Motivation: Systematic differences due to experimental features of microarray experiments are present in most large microarray data sets. Many different experimental features can cause biases including different sources of RNA, different production lots of microarrays or different microarray platforms. These systematic effects present a substantial hurdle to the analysis of microarray data.

Results: We present here a new method for the identification and adjustment of systematic biases that are present within microarray data sets. Our approach is based on modern statistical discrimination methods and is shown to be very effective in removing systematic biases present in a previously published breast tumor cDNA microarray data set. The new method of Distance Weighted Discrimination (DWD) is shown to be better than Support Vector Machines and Singular Value Decomposition for the adjustment of systematic microarray effects. In addition, it is shown to be of general use as a tool for the discrimination of systematic problems present in microarray data sets, including the merging of two breast tumor data sets collected on different microarray platforms.

Availability: Matlab software to perform DWD can be retrieved from https://genome.unc.edu/pubs/up/dwd/

Contact: marron@email.unc.edu; cperou@med.unc.edu

Supplementary information: The complete figures that represent the cluster diagrams in Figure 6 and other figures are available at https://genome.unc.edu/pubs/up/dwd/

1 INTRODUCTION

DNA microarrays are a powerful tool for the study of complex systems and are being applied to many questions in the biological sciences. In particular, the study of human tumors using patterns of gene expression have identified many expression differences that can predict important clinical properties like the propensity to relapse (van’t Veer et al., 2002) or the survival outlook for a patient (Sorlie et al., 2001).

However, a challenge of clinical sample studies is that systematic biases due to different handling procedures are often present. Microarray experiments are often performed over many months because sample collection is prospective, with most samples being assayed soon after they are collected. Additionally, samples tissue are collected and processed at different institutions and may be assayed using different microarray print batches or platforms or using different array hybridization protocols.

These systematic biases are manifested as differences in gene expression patterns when one set of microarrays is directly compared with a second set of microarrays. When using “supervised” statistical analyses, systematic biases show themselves as a subset of genes that tend to be more highly expressed in one set of microarrays versus another, and a consistent subset of genes that are lower in expression in one set versus the other. These biases can typically be identified using a cross-platform normalization method, which is based on linked gene sample clustering of the given datasets. The second is the introduction and description of several general validation measures that can be used to assess and compare cross-study normalization methods. The proposed normalization method is applied to three existing breast cancer datasets, and is compared to several competing normalization methods using the proposed validation measures.

Availability: The supplementary materials and XPN Matlab code are publicly available at website: https://genome.unc.edu/xpn

Contact: shabalin@email.unc.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

High-throughput gene-expression microarrays are currently being applied in a wide variety of biomedical problems. There are now several widely used, commercially available, microarray platforms that measure gene expression in related, but different, ways. No matter which technology is used, the evaluation of gene-expression experiments usually begins with statistical analyses that take a variety of forms, including exploratory analysis (such as clustering), classification and assessment of differential expression.

The increasing number and availability of large-scale gene-expression studies of human and other organisms provides strong motivation for cross-study analyses that combine existing and/or new datasets. In a cross-study analysis, the data, relevant test statistics or conclusions of several studies are combined. The simultaneous analysis of different studies of
Data Set Integration

Gene Expression (mRNAs)
- 25,000-75,000/sample

microRNAs
- 1500/sample

DNA Copy Number
- 1,000,000/sample

DNA Methylation
- 450,000/sample

Somatic Mutations
- 25,000/sample

Image of tissue and mutant class categories
Gene Expression (mRNAs)
300-500 Signatures/Modules

DNA Copy Number
1,000,000/sample

DNA Methylation
450,000/sample

Somatic Mutations
25,000/sample

Clinical Data
6 features/sample

Data Set Integration

Chris Fan
The Data Set was
1) 5 public microarray data sets combined (DWD combined)
2) Only patients used were those who received no systemic therapy (n=550)
3) Endpoint was relapse-free survival at 7 years
4) Used 323 genomic signatures/modules and 6 clinical variables
Gene Expression Signatures/Modules

A) Homogeneously expressed gene set = take median or mean expression for each patient

Examples = Proliferation gene set

B) Heterogeneously expressed gene list = 1st Principle Component of gene expression

Some gene lists are heterogeneously expressed in a dataset, for example, in this gene cluster. The samples in the left part has about 3/2 genes down regulated (green), 1/3 up regulated (red). We cannot take median or mean of the gene expression to represent this gene list, because the signal will cancel each other. In this case, we take 1st PC of the gene expression.

C) Predetermined Models = OncotypeDX, Mammaprint, PAM50 ROR-PT, and many more

D) Correlations to Centroids

Gene Expression data matrix converted to correlation values

Example = Score for correlation to each Subtype Centroid
550 patients

NKI (141) + UNC (33) + Loi (42) + Ivshina (137) + Wang (197)

Total 550 Samples

Randomly split into training (2/3) and testing sets (1/3)

Train (359 Samples) + Test (191 Samples)

Fan et al., BMC 2011 (PMID: 21214954)
## A. All Patients (N = 550)

<table>
<thead>
<tr>
<th></th>
<th>Combined model</th>
<th>Genomics model</th>
<th>Clinical model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard Ratio</td>
<td>P-Value</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>Training (N = 550)</td>
<td>6.99 &lt; 1.0e-22</td>
<td>6.46 &lt; 1.0e-22</td>
<td>5.54 3.4e-06</td>
</tr>
<tr>
<td>Testing (N = 191)</td>
<td>3.71 7.3e-07</td>
<td>3.92 1.1e-05</td>
<td>4.54 6.001</td>
</tr>
</tbody>
</table>

## B. ER-positive Patients (N = 395)

<table>
<thead>
<tr>
<th></th>
<th>Combined model</th>
<th>Genomics model</th>
<th>Clinical model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard Ratio</td>
<td>P-Value</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>Training (N = 395)</td>
<td>6.27 &lt; 1.0e-22</td>
<td>6.16 &lt; 1.0e-22</td>
<td>3.83 2.3e-06</td>
</tr>
<tr>
<td>Testing (N = 150)</td>
<td>2.58 6.5e-04</td>
<td>4.58 8.8e-04</td>
<td>6.16 0.0052</td>
</tr>
</tbody>
</table>

## C. ER-negative Patients (N = 155)

<table>
<thead>
<tr>
<th></th>
<th>Combined model</th>
<th>Genomics model</th>
<th>Clinical model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard Ratio</td>
<td>P-Value</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>Training (N = 100)</td>
<td>4.18 1.4e-0</td>
<td>6.66 7.9e-09</td>
<td>4.56 0.013</td>
</tr>
<tr>
<td>Testing (N = 66)</td>
<td>1.43 5.9e-03</td>
<td>1.41 0.11</td>
<td>1.29 0.73</td>
</tr>
</tbody>
</table>

## D. HER2-positive Patients (N = 110)

<table>
<thead>
<tr>
<th></th>
<th>Combined model</th>
<th>Genomics model</th>
<th>Clinical model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hazard Ratio</td>
<td>P-Value</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>Training (N = 75)</td>
<td>7.63 4.6e-8</td>
<td>7.63 8.1e-08</td>
<td>2.56 0.011</td>
</tr>
<tr>
<td>Testing (N = 37)</td>
<td>0.891 0.67</td>
<td>0.772 0.83</td>
<td>0.034 0.05</td>
</tr>
</tbody>
</table>

P-values are shown next to the respective hazard ratio values for each model.
<table>
<thead>
<tr>
<th>Gene</th>
<th>All Patients (n=550)</th>
<th>ER+ Patients (n=395)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGG_Cluster</td>
<td>98%</td>
<td>90%</td>
</tr>
<tr>
<td>E2F1_Repressed_by_Serum</td>
<td>81%</td>
<td>22%</td>
</tr>
<tr>
<td>Pcorr_NK170_Good</td>
<td>79%</td>
<td>89%</td>
</tr>
<tr>
<td>19p13_Amplicon</td>
<td>71%</td>
<td>21%</td>
</tr>
<tr>
<td>HS_Green19</td>
<td>62%</td>
<td>48%</td>
</tr>
<tr>
<td>MUnknown_28</td>
<td>56%</td>
<td>39%</td>
</tr>
<tr>
<td>MKRAS_amplicon</td>
<td>56%</td>
<td>39%</td>
</tr>
<tr>
<td>MHistone</td>
<td>46%</td>
<td>28%</td>
</tr>
<tr>
<td>HS_Green22</td>
<td>42%</td>
<td>28%</td>
</tr>
<tr>
<td>MM-Green23</td>
<td>38%</td>
<td>28%</td>
</tr>
<tr>
<td>Oncogenic_MYC</td>
<td>37%</td>
<td>28%</td>
</tr>
<tr>
<td>MM_Red21</td>
<td>33%</td>
<td>21%</td>
</tr>
<tr>
<td>E2F1_NOT_Repressed_by_Serum</td>
<td>28%</td>
<td>28%</td>
</tr>
<tr>
<td>16.13_Amplicon</td>
<td>27%</td>
<td>16%</td>
</tr>
<tr>
<td>Unknown_2</td>
<td>27%</td>
<td>16%</td>
</tr>
<tr>
<td>HS_Red16</td>
<td>21%</td>
<td>16%</td>
</tr>
<tr>
<td>MUnknown_20</td>
<td>20%</td>
<td>16%</td>
</tr>
<tr>
<td>MM_Red18</td>
<td>19%</td>
<td>16%</td>
</tr>
<tr>
<td>Fibrinogen_Cluster</td>
<td>14%</td>
<td>16%</td>
</tr>
<tr>
<td>MUnknown_15</td>
<td>12%</td>
<td>11%</td>
</tr>
<tr>
<td>MUnknown_21</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>MUnknown_1</td>
<td>12%</td>
<td>11%</td>
</tr>
<tr>
<td>ROR_S</td>
<td>20%</td>
<td>21%</td>
</tr>
<tr>
<td>MUnknown_34</td>
<td>24%</td>
<td>21%</td>
</tr>
<tr>
<td>Oncogenic_BCAT</td>
<td>27%</td>
<td>21%</td>
</tr>
<tr>
<td>CD44+PROCR+-vs-CD24+-Downregulated</td>
<td>28%</td>
<td>12%</td>
</tr>
<tr>
<td>Unknown_9</td>
<td>31%</td>
<td>23%</td>
</tr>
<tr>
<td>MUnknown_30</td>
<td>32%</td>
<td>23%</td>
</tr>
<tr>
<td>MM_Red23</td>
<td>36%</td>
<td>23%</td>
</tr>
<tr>
<td>MNOrth4</td>
<td>43%</td>
<td>23%</td>
</tr>
<tr>
<td>GHI_RS</td>
<td>44%</td>
<td>23%</td>
</tr>
<tr>
<td>VEGF_13genes</td>
<td>47%</td>
<td>28%</td>
</tr>
<tr>
<td>MM_Red10</td>
<td>50%</td>
<td>26%</td>
</tr>
<tr>
<td>ADM_S100A10_A110NDGR1_Cluster</td>
<td>56%</td>
<td>25%</td>
</tr>
<tr>
<td>HER2_Amplicon</td>
<td>71%</td>
<td>25%</td>
</tr>
<tr>
<td>Scorr_Her2</td>
<td>74%</td>
<td>22%</td>
</tr>
<tr>
<td>Tumor Size</td>
<td>75%</td>
<td>21%</td>
</tr>
<tr>
<td>16q24x</td>
<td>89%</td>
<td>21%</td>
</tr>
<tr>
<td>Histological Grade</td>
<td>100%</td>
<td>94%</td>
</tr>
<tr>
<td>Rotterdam_76_Gene</td>
<td>100%</td>
<td>99%</td>
</tr>
</tbody>
</table>

Fan et al., BMC 2011 (PMID: 21214954)
All Patients (n=550) vs. ER+ Patients (n=395)

- Clinical
- Genomics.323
- Combined.323
- Genomics.319
- Combined.319
- ROR_S
- ROR_C (Subtype + Clinical)
- Rotterdam-76
- Rotterdam-76 + Clinical
- OncoTypeDX RS
- OncoTypeDX RS + Clinical
- NKI-70
- NKI-70 + Clinical

Mean Concordance Index (C-Index)

Fan et al., BMC 2011 (PMID: 21214954)
Data Set Integration using Clinical, RNAseq, and DNAseq (exomes)

Gene Expression (mRNAs) 300-500 Signatures/Modules

Clinical Data 6 features/sample

Proteins 200/sample

DNA Copy Number 1,000,000/sample

DNA Methylation 450,000/sample

Somatic Mutations 25,000/sample

Gene Expression (mRNAs)

Clinical Data

Proteins

DNA Copy Number

DNA Methylation

Somatic Mutations
## Estimated new cases*

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>238,590</td>
<td>28%</td>
</tr>
<tr>
<td>Lung &amp; bronchus</td>
<td>118,080</td>
<td>14%</td>
</tr>
<tr>
<td>Colorectum</td>
<td>73,680</td>
<td>9%</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>54,610</td>
<td>6%</td>
</tr>
<tr>
<td>Melanoma of the skin</td>
<td>45,060</td>
<td>5%</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>40,430</td>
<td>5%</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>37,600</td>
<td>4%</td>
</tr>
<tr>
<td>Oral cavity &amp; pharynx</td>
<td>29,620</td>
<td>3%</td>
</tr>
<tr>
<td>Leukemia</td>
<td>27,880</td>
<td>3%</td>
</tr>
<tr>
<td>Pancreas</td>
<td>22,740</td>
<td>3%</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>854,790</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

## Estimated deaths

<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung &amp; bronchus</td>
<td>87,260</td>
<td>28%</td>
</tr>
<tr>
<td>Prostate</td>
<td>29,720</td>
<td>10%</td>
</tr>
<tr>
<td>Colorectum</td>
<td>26,300</td>
<td>9%</td>
</tr>
<tr>
<td>Pancreas</td>
<td>19,480</td>
<td>6%</td>
</tr>
<tr>
<td>Liver &amp; intrahepatic bile duct</td>
<td>14,890</td>
<td>5%</td>
</tr>
<tr>
<td>Leukemia</td>
<td>13,660</td>
<td>4%</td>
</tr>
<tr>
<td>Esophagus</td>
<td>12,220</td>
<td>4%</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>10,820</td>
<td>4%</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>10,590</td>
<td>3%</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>8,780</td>
<td>3%</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>306,920</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Lung cancer subtypes

- Non-small-cell lung carcinoma (NSCLC, 85%)
  - adenocarcinoma (LAD, 40%)
    - Most cases of adenocarcinoma are associated with smoking. Adenocarcinoma is the most common form of lung cancer for never smokers
      - Bronchioid
      - Magnoid
      - Squamoid
  - squamous-cell lung carcinoma (SCC, 30%)
    - Primitive
    - Classical
    - Secretory
    - Basal (worst outcome)
  - large-cell lung carcinoma (LC, 9%)

- Small-cell lung carcinoma (SCLC, 10%-15%)
  - These cancers grow quickly and spread early in the course of the disease. This type of lung cancer is strongly associated with smoking.

- Other Subtypes
  - glandular tumors
  - carcinoid tumors
  - adenosquamous carcinoma
  - sarcomatoid carcinoma
  - undifferentiated carcinomas

We selected Adenocarcinomas because of its sample size and clinical features predicted outcomes
356 TCGA LUAD patients with OS data and OS time > 20 days
Split data into Training and Testing Sets

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>122 (60.4%)</th>
<th>80 (39.6%)</th>
<th>0.9134</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>97 (61.4%)</td>
<td>61 (38.6%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathologic_T</th>
<th>T1</th>
<th>67 (61.5%)</th>
<th>42 (38.5%)</th>
<th>0.4676</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>123 (60.9%)</td>
<td>79 (39.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>21 (67.7%)</td>
<td>10 (32.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>7 (43.8%)</td>
<td>9 (56.3%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pathologic_N</th>
<th>N0</th>
<th>135 (60.0%)</th>
<th>90 (40.0%)</th>
<th>0.5474</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1</td>
<td>45 (66.2%)</td>
<td>23 (33.8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>32 (57.1%)</td>
<td>24 (42.9%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Total | 219 (60.8%) | 141 (39.2%) |

A Fast Algorithm for Balanced Sampling

Guillaume Chauvet¹ and Yves Tillé²

¹ Laboratoire de Statistique d’Enquête, CREST - ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz, France.
² Statistics Group, University of Neuchâtel, Espace de l’Europe 4, Case postale 827, 2002 Neuchâtel, Switzerland.
Split data into Training and Testing Sets

**pathologic_T**

- **Train**
  - Months: 1 2 3 4
  - Overall Survival (Probability):
    - 1: 15/109
    - 2: 61/202
    - 3: 9/31
    - 4: 9/16
  - log rank p = 0.00116

- **Test**
  - Months: 1 2 3 4
  - Overall Survival (Probability):
    - 1: 6/42
    - 2: 23/79
    - 3: 2/10
    - 4: 6/9
  - log rank p = 0.0165

**pathologic_N**

- **Train**
  - Months: 0 1 2
  - Overall Survival (Probability):
    - 0: 39/225
    - 1: 31/68
    - 2: 22/56
  - log rank p = 3.4e-05

- **Test**
  - Months: 0 1 2
  - Overall Survival (Probability):
    - 0: 16/90
    - 1: 11/23
    - 2: 10/24
  - log rank p = 0.00187
Cluster of 390 Gene Expression Signatures in 356 Lung Adenocarcinoma Patients
(200 signatures from Perou Lab and 190 from other publications,
see Fan et al., BMC Medical Genomics, 2011 PMID:21214954)
356 LUAD Patients, 1371 features (3 data types)

- 807 DNA Copy Number Altered Regions (segments)
- 174 Mutations
- 390 Gene Expression Signatures

Elastic Net Modeling using Multiple Data Types on TCGA Lung Adenocarcinoma Samples done by Chris Fan of the Lineberger Bioinformatics Group

Fan and Perou 2016, unpublished data, do not distribute
We used Overall Survival at 5 years as the “Supervising” prognostic endpoint

COX Proportional Hazards Models built using 5-fold Cross Validation (CV) on a training data set (60% used for training)

Predict onto testing data (40% used as test set)

Report training and testing sets Kaplan-Meier survival plot p-values and Harrell's C-Index values

Fan and Perou 2016, unpublished data, do not distribute
ELASTIC NET is a modeling approach that can be used to perform both feature selection (from multiple data types) and parameter estimation. It is a hybrid of Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO) Regression. Like the LASSO, ELASTIC NET performs automatic feature selection and shrinkage to produce sparse models with high prediction accuracy.

LASSO sometimes fails to do grouped feature selection, and it tends to select one feature from a group of correlated features and ignore the others. ELASTIC NET does not have this limitation, and seems to strike a good balance between selecting just one correlated feature versus selecting all correlated features (Hastie, [http://www-stat.stanford.edu/~hastie/TALKS/glmnet.pdf](http://www-stat.stanford.edu/~hastie/TALKS/glmnet.pdf)).

Fan and Perou 2016, unpublished data, do not distribute
mRNA Signatures only

<table>
<thead>
<tr>
<th>Study Name</th>
<th>Median</th>
<th>PMID</th>
<th>C-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNC_MPYMT_NEU_Cluster_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.151608</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_MM_Green12_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0436784</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_MUnknown_20_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0359065</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_IGG_Cluster_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0186588</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_MM_Red17_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0165798</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_VEGF_13genes_Median_BMC.Med.2009_PMID.19291283</td>
<td>0.0130258</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_HS_Green19_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0253536</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_13q14_Amplicon_Median_BMC.Med.Genomics.2011_PMID.21214954</td>
<td>0.0391103</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_Scorr_Basal_Correlation_JCO.2009_PMID.19204204</td>
<td>0.0514695</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>UNC_MM_p53null.Basal_MedianFrom.Adam.Pfefferle_Pending_PMID.24220145</td>
<td>0.0877532</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Extensive_Residual_Diesase_ER54_Median_JAMA.2011_PMID.21558518</td>
<td>0.1121707</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Train**

- C-index = 0.729
- Low: 15/109 (P = 7.25e-07)
- High: 46/109

**Test**

- C-index = 0.662
- Low: 14/69 (P = 0.0185)
- High: 24/69

---

Fan and Perou 2016, unpublished data, do not distribute
DNA Copy Number Changes only

Train

C-index = 0.719

Test

C-index = 0.527

Overall Survival (Probability)

low 17/109
high 44/109
P = 7.1e-06

low 16/69
high 22/69
log rank p=0.733

overall survival

Train

Test

log rank

C-index

DNA.Cytoband_19p13.12 -0.13738
DNA.Cytoband_19q11 -0.07261
DNA.Cytoband_6q11.1 -0.05323
DNA.Cytoband_19p12 -0.039
DNA.Cytoband_6q12 -0.03579
DNA.Cytoband_1p36.11 -0.01076
DNA.Cytoband_10p12.33 -0.00808
DNA.Cytoband_10p15.3 -0.00275
DNA.Cytoband_19p13.3 -0.00263
DNA.Cytoband_22q13.32 -0.003424
DNA.Cytoband_22q13.33 -0.003443
DNA.Cytoband_16q24.3 -0.005053
DNA.Cytoband_8p11.21 -0.010792
DNA.Cytoband_3q26.31 -0.013839
DNA.Cytoband_16q24.2 -0.044561
DNA.Cytoband_14q21.2 -0.07531
DNA.Cytoband_Xp11.1 -0.089138
DNA.Cytoband_22q13.2 -0.115788

Fan and Perou 2016, unpublished data, do not distribute
DNA Mutations (Exomes only)

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation_FAM75D1</td>
<td>-0.8966703</td>
</tr>
<tr>
<td>Mutation_ZAN</td>
<td>-0.7062429</td>
</tr>
<tr>
<td>Mutation_MUC5B</td>
<td>-0.6747349</td>
</tr>
<tr>
<td>Mutation_LRRC7</td>
<td>-0.4113941</td>
</tr>
<tr>
<td>Mutation_TSHZ3</td>
<td>-0.3721812</td>
</tr>
<tr>
<td>Mutation_AHNK2</td>
<td>-0.3242386</td>
</tr>
<tr>
<td>Mutation_ANKRD20A9P</td>
<td>-0.2811364</td>
</tr>
<tr>
<td>Mutation_CUBN</td>
<td>-0.2718021</td>
</tr>
<tr>
<td>Mutation_HMCN1</td>
<td>-0.2343693</td>
</tr>
<tr>
<td>Mutation_RYR2</td>
<td>-0.2308531</td>
</tr>
<tr>
<td>Mutation_EPHA5</td>
<td>-0.1921632</td>
</tr>
<tr>
<td>Mutation_CNTNAP5</td>
<td>-0.1647605</td>
</tr>
<tr>
<td>Mutation_LAMA2</td>
<td>-0.1629955</td>
</tr>
<tr>
<td>Mutation_SORCS1</td>
<td>-0.1595584</td>
</tr>
<tr>
<td>Mutation_CSMD3</td>
<td>-0.130134</td>
</tr>
<tr>
<td>Mutation_PCDH10</td>
<td>-0.1001778</td>
</tr>
<tr>
<td>Mutation_ZNF208</td>
<td>-0.0930673</td>
</tr>
<tr>
<td>Mutation_TPTE</td>
<td>-0.0702816</td>
</tr>
<tr>
<td>Mutation_BC080605</td>
<td>-0.0304059</td>
</tr>
<tr>
<td>Mutation_DNAH7</td>
<td>-0.028821</td>
</tr>
<tr>
<td>Mutation_STK11</td>
<td>0.00957196</td>
</tr>
<tr>
<td>Mutation_EGFR</td>
<td>0.01215757</td>
</tr>
<tr>
<td>Mutation_TRPS1</td>
<td>0.01481218</td>
</tr>
<tr>
<td>Mutation_SSPO</td>
<td>0.09275657</td>
</tr>
<tr>
<td>Mutation_USH2A</td>
<td>0.10742406</td>
</tr>
<tr>
<td>Mutation_PRUNE2</td>
<td>0.13405142</td>
</tr>
<tr>
<td>Mutation_GPR98</td>
<td>0.16798704</td>
</tr>
<tr>
<td>Mutation_KEAP1</td>
<td>0.24747142</td>
</tr>
<tr>
<td>Mutation_TP53</td>
<td>0.38780379</td>
</tr>
<tr>
<td>Mutation_PCDHGC5</td>
<td>0.4229068</td>
</tr>
<tr>
<td>Mutation_ANKRD30A</td>
<td>0.49702014</td>
</tr>
<tr>
<td>Mutation_COL6A3</td>
<td>0.6345502</td>
</tr>
</tbody>
</table>

Fan and Perou 2016, unpublished data, do not distribute
mRNA Signatures
+ DNA Copy Number
+ DNA mutations

Fan and Perou 2016, unpublished data, do not distribute
**Clinical Data only**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender_female_0</td>
<td>-0.27192</td>
</tr>
<tr>
<td>Age at initial pathologic diagnosis</td>
<td>0.019087</td>
</tr>
<tr>
<td>Pathologic N</td>
<td>0.528502</td>
</tr>
<tr>
<td>Pathologic T</td>
<td>0.643398</td>
</tr>
</tbody>
</table>

Train

C-index = 0.713

Test

C-index = 0.643

Fan and Perou 2016, unpublished data, do not distribute
mRNA Signatures + DNA Copy Number + DNA mutations + Clinical Data

Fan and Perou 2016, unpublished data, do not distribute
1. Identification of patterns/structure in the data using unsupervised and supervised learning methods are key approaches for Big Data Mining. Building formal classifiers based upon this structure provides an objective measure for assessing the value of the selected features.

2. Methods like Elastic Net and LASSO can be used to develop predictors of outcomes, or response (binary variable), using single platform and/or multi-platform integrated data.

3. Integration of data across technology platforms for biological discovery and assay development is becoming increasingly important.

4. Identification (and correction) of technical batch effects, or other systematic features present within large data sets, is a must for the effective mining of Big Data Resources.
Take the “classification list” data, and make a hierarchical cluster.

1) I recommend using “Cluster 3.0” and under “adjust data” select log2 transform and median center the genes.

2) Next select the “hierarchical” tab and select cluster (both genes and arrays), similarity metric “centered” and cluster method “centroid linkage”.

3) The data can then be viewed and explored using “Java Treeview”. Look for and find the “proliferation cluster”, which was showed earlier in this presentation.