nCounter Assay

Automated Process

Hybridize CodeSet 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

nCounter Prep Station

Hybridize Reporter to RNA

Remove excess reporters

Bind reporter to surface

Immobilize and align reporter

Image surface

Count codes

Hybridized mRNA

slides courtesy of Sean Ferree, Nanostring Technologies
nCounter Assay

Automated Process

nCounter Prep Station

1. Hybridize Reporter to RNA
2. Remove excess reporters
3. Bind reporter to surface
4. Immobilize and align reporter
5. Image surface
6. Count codes

Surface of cartridge is coated with streptavidin

slides courtesy of Sean Ferree, Nanostring Technologies
nCounter Assay

Automated Process

nCounter Prep Station

1. Hybridize Reporter to RNA
2. Remove excess reporters
3. Bind reporter to surface
4. Immobilize and align reporter
5. Image surface
6. Count codes

slides courtesy of Sean Ferree, Nanostring Technologies
nCounter Assay

Automated Process

1. Hybridize Reporter to RNA
2. Remove excess reporters
3. Bind reporter to surface
4. Immobilize and align reporter
5. Image surface
6. Count codes

One coded reporter = 1 mRNA = 1 gene

Slides courtesy of Sean Ferree, Nanostring Technologies
Combined Analysis: Clinical Validation in Over 2,400 Patients

- Prospectively defined analysis of two registration-quality databases with ≥10-yr median follow-up in postmenopausal women with ER+ ESBC treated with endocrine therapy alone
  - **Primary objective:** Validate published observations that the ROR score provides additional prognostic information above standard clinical variables for DRFS at 10 yrs
    - Primary Analysis: All patients
    - Secondary Analysis: Node -/+ , HER2-negative patients
  - **Secondary objective:** Validate observations that Luminal A and Luminal B patients have statistically significantly different DRFS at 10 years

<table>
<thead>
<tr>
<th>transATAC</th>
<th>ABCSG-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>• N = 1,007</td>
<td>• N = 1,478</td>
</tr>
<tr>
<td>• Patients: Postmenopausal women with ER/PR+ early-stage breast cancer treated with endocrine therapy alone</td>
<td>• Patients: Postmenopausal women with ER/PR+ early-stage breast cancer treated with endocrine therapy alone</td>
</tr>
<tr>
<td>• Design: Prospective, retrospective</td>
<td>• Design: Prospective, retrospective</td>
</tr>
<tr>
<td>• Output: Prognosis</td>
<td>• Output: Prognosis</td>
</tr>
</tbody>
</table>

---

ROR Defined Risk Groups have statistically significant different outcomes at 10 years – Node Negative Patients

<table>
<thead>
<tr>
<th>Risk group</th>
<th>Patients (%)</th>
<th>Number of Events through 10 years</th>
<th>Estimated percentage of risk at 10 years (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>166 (10.0)</td>
<td>31</td>
<td>20.3 (14.7 - 27.7)</td>
</tr>
<tr>
<td>intermediate</td>
<td>388 (23.5)</td>
<td>52</td>
<td>15.0 (11.6 - 19.2)</td>
</tr>
<tr>
<td>low</td>
<td>1100 (66.5)</td>
<td>50</td>
<td>4.9 (3.7 - 6.4)</td>
</tr>
</tbody>
</table>

Patients at risk
- high: 166, 164, 156, 147, 141, 132, 124, 120, 114, 88, 73, 48, 34, 25, 13, 7
- low: 1100, 1089, 1067, 1051, 1039, 1024, 996, 974, 933, 812, 562, 265, 182, 115, 72, 34

Follow-up time (years)
Prediction of late distant recurrence after 5 years of endocrine treatment: a combined analysis of patients from the Austrian breast and colorectal cancer study group 8 and anastrozole, tamoxifen alone or in combination randomized trials using the PAM50 risk of recurrence score.

Sestak et al., JCO 2014 (PMID: 25332252)

**Fig 2.** Annual hazard rate curve for all patients according to risk of recurrence groups.
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**
- **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

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

Genomic DNA

Introns

Exons

Fragmentation (100-300 bp), denaturation

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

Alignment
- MapSplice
- STAR
- BWA

BAM
- COUNTS (exons, transcript level)
- SPLICING (novel transcripts & splice junctions)
- MUTATIONS
- FUSIONS

SRF
- FASTQ

UNC Tools
- RSEM data matrix

MapSplice
- Text
- MAF

GEO: Gene Expression Omnibus
- UNC Microarray Database

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)

cor=0.994
Agilent 244K microarray versus mRNA-seq
(Log2 ratio created for mRNA-seq using Common Reference)

Correlation of log2(Ratio) All (13
Microarray || 9830-020018B-244Kv5
breast tumor || 9830-020018B-MG | 100902_UN
cor = 0.872

Correlation of log2(Ratio) Intrins
Microarray || 9830-020018B-MG | 100902_UN
breast tumor || 9830-020018B-244Kv5
cor = 0.9
mRNA-seq data was “platform” normalized to Agilent data
Adjustment of systematic microarray data biases

Monica Benito1, Joel Parker2, Quan Du2, Junyuan Wu2, Dong Xiang3, Charles M. Perou2,5,6,* and J. S. Marron5,6,*

1Department of Statistics and Econometrics, University of Carlos III, Madrid, Spain.
2Uniberg Comprehensive Cancer Center, 3Department of Genetics and 4Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599-7284, USA
5Department of Molecular Medicine, Karolinska Institute, S-171 77 Stockholm, Sweden
6Department of Statistics, University of North Carolina, Chapel Hill, NC
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 completed on different microarray platforms.

Availability: Matlab software to perform DWD can be retrieved from https://genomics.unc.edu/putisp/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://genomics.unc.edu/putisp/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 Veen et al., 2002) or the survival outcome for a patient (Serlie 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 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 consequent subset of genes that are lower in expression in one set versus the other. These biases can typically be identified by the use of ‘root mean squares’ or ‘mean squares’ statistics, which can detect differences between arrays, but which do not adjust for differences between experiments.

Gene expression
Merging two gene-expression studies via cross-platform normalization

Andrey A. Shabalin1, Hakon Tjelmeland2, Cheng Fan3, Charles M. Perou4,5,6 and Andrew B. Nobel1

1Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, USA
2Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
3Uniberg Comprehensive Cancer Center, University of North Carolina at Chapel Hill, USA
4Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill
5Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, USA

Received on November 28, 2007; revised on February 7, 2008; accepted on March 1, 2008

Advance Access publication March 5, 2008

Associate Editor: David Rocke

ABSTRACT
Motivation: Gene-expression microarrays are currently being applied in a variety of biomedical applications. This article considers the problem of how to merge datasets arising from different gene-expression studies of a common organism and phenotype. Of particular interest is how to merge data from different technological platforms.

Results: The article makes two contributions to the problem. The first is a simple cross-platform normalization method, which is based on linking gene/sample clustering of the given datasets. The second is the introduction of several general validation measures that can be used to assess and compare cross-platform normalization methods. The proposed normalization method is applied to three existing breast cancer datasets, and 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 assessments 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
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

<table>
<thead>
<tr>
<th>Modules Source</th>
<th>Methods</th>
<th>#Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Gene lists</td>
<td>Median</td>
<td>13</td>
</tr>
<tr>
<td>Bi-Cluster identified from Mouse</td>
<td>Median</td>
<td>50</td>
</tr>
<tr>
<td>Bi-Cluster identified from Human</td>
<td>Median</td>
<td>50</td>
</tr>
<tr>
<td>Unsupervised Cluster from Human</td>
<td>Median</td>
<td>52</td>
</tr>
<tr>
<td>Unsupervised Cluster from Mouse</td>
<td>Median</td>
<td>56</td>
</tr>
<tr>
<td>Published Gene lists</td>
<td>First PCA</td>
<td>77</td>
</tr>
<tr>
<td>Published Gene lists</td>
<td>Correlations</td>
<td>22</td>
</tr>
<tr>
<td>Published Models</td>
<td>Hazard Ratio</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total Modules</strong></td>
<td></td>
<td><strong>323</strong></td>
</tr>
</tbody>
</table>
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
Fan et al., BMC 2011 (PMID: 21214954)

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)
### All Patients (n=550)

<table>
<thead>
<tr>
<th>Genes/Clusters</th>
<th>FREQ %</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGG_Cluster</td>
<td>98</td>
<td></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>3.4</td>
</tr>
<tr>
<td>19p13_Amplicon</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>HS_Green19</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>MUnknown_28</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>MKRAS_amplicon</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>MHistone</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>HS_Green22</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>MM-Green23</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Oncogenic_MYC</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>MM_Red21</td>
<td>33</td>
<td>10</td>
</tr>
<tr>
<td>E2F1_NOT_Repressed_by_Serum</td>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>16.13_Amplicon</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Unknown_2</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>HS_Red16</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>MUnknown_20</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>MM_Red18</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Fibrinogen_Cluster</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>MUnknown_15</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>MUnknown_21</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>MUnknown_1</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>ROR_S</td>
<td>20</td>
<td>46</td>
</tr>
<tr>
<td>MUnknown_34</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Oncogenic_BCAT</td>
<td>27</td>
<td>10</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></td>
</tr>
<tr>
<td>MUnknown_30</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>MM_Red23</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>MNOTch4</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>GHI_RS</td>
<td>44</td>
<td>5</td>
</tr>
<tr>
<td>VEGF_13genes</td>
<td>47</td>
<td>26</td>
</tr>
<tr>
<td>MM_Red10</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>ADM_S100A10_A110NDGR1_Cluster</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>HER2_Amplicon</td>
<td>71</td>
<td></td>
</tr>
<tr>
<td>Scoor_Her2</td>
<td>74</td>
<td>50</td>
</tr>
<tr>
<td>Tumor Size</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>16q24</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>Histological Grade</td>
<td>100</td>
<td>47</td>
</tr>
<tr>
<td>Rotterdam_76_Gene</td>
<td>100</td>
<td>47</td>
</tr>
</tbody>
</table>

### ER+ Patients (n=395)

<table>
<thead>
<tr>
<th>Genes/Clusters</th>
<th>FREQ %</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGG_Cluster</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>Pcorr_NK170_Good</td>
<td>89</td>
<td>3.4</td>
</tr>
<tr>
<td>MUnknown_28</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>MM_Red21</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>MHistone</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>HS_Green19</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>MM-Green23</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>Oncogenic_MYC</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>MUnknown_20</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>E2F1_NOT_Repressed_by_Serum</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>1p36_Amplicon</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Oncogenic_MYC</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>$100A9_A8</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Unknown_2</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>19p13_Amplicon</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>HS_Green22</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>IDGR</td>
<td>7</td>
<td>63</td>
</tr>
<tr>
<td>CD44+PROCR+vs-CD24+Downregulated</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>Scoor_Her2</td>
<td>23</td>
<td>50</td>
</tr>
<tr>
<td>ROR_S</td>
<td>23</td>
<td>46</td>
</tr>
<tr>
<td>Bone_Metastasis_Underexpressed</td>
<td>24</td>
<td>58</td>
</tr>
<tr>
<td>VEGF_13genes</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>MUnknown_28</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Glycolysis_Signature</td>
<td>29</td>
<td>26</td>
</tr>
<tr>
<td>MUnknown_1</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>HER1_Cluster2</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>LKB1</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>GHI_RS</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>MNOTch4</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>HER2_Amplicon</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>ADM_S100A10_A110NDGR1_Cluster</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>MM_Red10</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>10q24</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Tumor Size</td>
<td>61</td>
<td></td>
</tr>
<tr>
<td>MUnknown_30</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Histological Grade</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Rotterdam_76_Gene</td>
<td>99</td>
<td>47</td>
</tr>
</tbody>
</table>

Fan et al., BMC 2011 (PMID: 21214954)
All Patients (n=550)  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)
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
## Estimated new cases*

<table>
<thead>
<tr>
<th>Site</th>
<th>Males</th>
<th>Females</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate</td>
<td>238,590</td>
<td>28%</td>
<td>Breast</td>
<td>232,340</td>
</tr>
<tr>
<td>Lung &amp; bronchus</td>
<td><strong>118,080</strong></td>
<td><strong>14%</strong></td>
<td>Lung &amp; bronchus</td>
<td><strong>110,110</strong></td>
</tr>
<tr>
<td>Colorectum</td>
<td>73,680</td>
<td>9%</td>
<td>Colorectum</td>
<td>69,140</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>54,610</td>
<td>6%</td>
<td>Uterine corpus</td>
<td>49,560</td>
</tr>
<tr>
<td>Melanoma of the skin</td>
<td>45,060</td>
<td>5%</td>
<td>Thyroid</td>
<td>45,310</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>40,430</td>
<td>5%</td>
<td>Non-Hodgkin lymphoma</td>
<td>32,140</td>
</tr>
<tr>
<td>Non-Hodgkin lymphoma</td>
<td>37,600</td>
<td>4%</td>
<td>Melanoma of the skin</td>
<td>31,630</td>
</tr>
<tr>
<td>Oral cavity &amp; pharynx</td>
<td>29,620</td>
<td>3%</td>
<td>Kidney &amp; renal pelvis</td>
<td>24,720</td>
</tr>
<tr>
<td>Leukemia</td>
<td>27,880</td>
<td>3%</td>
<td>Pancreas</td>
<td>22,480</td>
</tr>
<tr>
<td>Pancreas</td>
<td>22,740</td>
<td>3%</td>
<td>Ovary</td>
<td>22,240</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>854,790</strong></td>
<td><strong>100%</strong></td>
<td><strong>All Sites</strong></td>
<td><strong>805,500</strong></td>
</tr>
</tbody>
</table>

## Estimated deaths

<table>
<thead>
<tr>
<th>Site</th>
<th>Males</th>
<th>Females</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung &amp; bronchus</td>
<td><strong>87,260</strong></td>
<td><strong>28%</strong></td>
<td>Lung &amp; bronchus</td>
<td><strong>72,220</strong></td>
</tr>
<tr>
<td>Prostate</td>
<td>29,720</td>
<td>10%</td>
<td>Breast</td>
<td>39,620</td>
</tr>
<tr>
<td>Colorectum</td>
<td>26,300</td>
<td>9%</td>
<td>Colorectum</td>
<td>24,530</td>
</tr>
<tr>
<td>Pancreas</td>
<td>19,480</td>
<td>6%</td>
<td>Pancreas</td>
<td>18,980</td>
</tr>
<tr>
<td>Liver &amp; intrahepatic bile duct</td>
<td>14,890</td>
<td>5%</td>
<td>Ovary</td>
<td>14,030</td>
</tr>
<tr>
<td>Leukemia</td>
<td>13,660</td>
<td>4%</td>
<td>Leukemia</td>
<td>10,060</td>
</tr>
<tr>
<td>Esophagus</td>
<td>12,220</td>
<td>4%</td>
<td>Non-Hodgkin lymphoma</td>
<td>8,430</td>
</tr>
<tr>
<td>Urinary bladder</td>
<td>10,820</td>
<td>4%</td>
<td>Uterine corpus</td>
<td>8,190</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma</td>
<td>10,590</td>
<td>3%</td>
<td>Liver &amp; intrahepatic bile duct</td>
<td>6,780</td>
</tr>
<tr>
<td>Kidney &amp; renal pelvis</td>
<td>8,780</td>
<td>3%</td>
<td>Brain &amp; other nervous system</td>
<td>6,150</td>
</tr>
<tr>
<td><strong>All Sites</strong></td>
<td><strong>306,920</strong></td>
<td><strong>100%</strong></td>
<td><strong>All Sites</strong></td>
<td><strong>271,430</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
<table>
<thead>
<tr>
<th>Available Cancer Types</th>
<th># Cases Shipped by BCR</th>
<th># Cases with Data</th>
<th>Date Last Updated (mm/dd/yy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung adenocarcinoma [LUAD]</td>
<td>521</td>
<td>521</td>
<td>01/30/15</td>
</tr>
<tr>
<td>Lung squamous cell carcinoma [LUSC]</td>
<td>510</td>
<td>504</td>
<td>01/26/15</td>
</tr>
<tr>
<td>Breast invasive carcinoma [BRCA]</td>
<td>1100</td>
<td>1098</td>
<td>01/28/15</td>
</tr>
</tbody>
</table>

356 TCGA LUAD patients with OS data and OS time > 20 days
Split data into Training and Testing Sets

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
<th>Fisher's Test P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>122 (60.4%)</td>
<td>80 (39.6%)</td>
<td>0.9134</td>
</tr>
<tr>
<td>Male</td>
<td>97 (61.4%)</td>
<td>61 (38.6%)</td>
<td></td>
</tr>
<tr>
<td><strong>Pathologic_T</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>67 (61.5%)</td>
<td>42 (38.5%)</td>
<td>0.4676</td>
</tr>
<tr>
<td>T2</td>
<td>123 (60.9%)</td>
<td>79 (39.1%)</td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>21 (67.7%)</td>
<td>10 (32.3%)</td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>7 (43.8%)</td>
<td>9 (56.3%)</td>
<td></td>
</tr>
<tr>
<td><strong>pathologic_N</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N0</td>
<td>135 (60.0%)</td>
<td>90 (40.0%)</td>
<td>0.5474</td>
</tr>
<tr>
<td>N1</td>
<td>45 (66.2%)</td>
<td>23 (33.8%)</td>
<td></td>
</tr>
<tr>
<td>N2</td>
<td>32 (57.1%)</td>
<td>24 (42.9%)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>219 (60.8%)</td>
<td>141 (39.2%)</td>
<td></td>
</tr>
</tbody>
</table>

T.test P = 0.862

Overall Survival (Probabil): Test 37/141
Train 57/219

log rank p = 0.869

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

- **Test**
  - Overall Survival (Probability)
  - Log rank p = 0.0115
  - Months: 1 9/67, 2 38/123, 3 7/21, 4 3/7

**pathologic_N**

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

- **Test**
  - Overall Survival (Probability)
  - Log rank p = 0.00131
  - Months: 0 16/90, 1 11/23, 2 10/24
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)
Elastic Net Modeling using Multiple Data Types on TCGA Lung Adenocarcinoma Samples
done by Chris Fan of the Lineberger Bioinformatics Group

356 LUAD Patients, 1371 features (3 data types)

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

Fan and Perou 2017, 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 2017, 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).
mRNA Signatures only

### Raw Text

**Train**

- **Overall Survival (Probability)**
- **C-index = 0.729**
- **low 15/109**
- **high 46/109**
- **P = 7.25e-07**

**Test**

- **Overall Survival (Probability)**
- **C-index = 0.662**
- **low 14/69**
- **high 24/69**
- **P = 0.0185**

---

**Extensive Residual Disease ER54 Median JAMA.2011_PMID.21558518 0.1121707**

**UNC_Basal_MedianFrom.Adam.Pfefferle_Pending_PMID.24220145 0.0877532**

**UNC_MM_cluster_20_Median_BMC.Med.Genomics.2011_PMID.21214954 0.0359065**

**UNC_MM_cluster_17_Median_BMC.Med.Genomics.2011_PMID.21214954 0.0165798**

**UNC_VEGF_13genes_Median_BMC.Med.2009_PMID.19291283 0.0130258**

**UNC_HS_Green19_Median_BMC.Med.Genomics.2011_PMID.21214954 0.0253536**

**UNC_13q14_Amplicon_cluster_Median_BMC.Med.Genomics.2011_PMID.21214954 0.0391103**

**UNC_Scorr_Basal_Correlation_JCO.2009_PMID.19204204 0.0514695**

**UNC_MM_p53null_Basal_MedianFrom.Adam.Pfefferle_Pending_PMID.24220145 0.0877532**

---

Fan and Perou 2017, unpublished data, do not distribute
DNA Copy Number Changes only

Train

Overall Survival (Probability)

C-index = 0.719

Test

Overall Survival (Probability)

C-index = 0.527

low 17/109

high 44/109

P = 7.1e-06

low 16/69

high 22/69

log rank p=0.733

Fan and Perou 2017, unpublished data, do not distribute
DNA Mutations (Exomes only)

Fan and Perou 2017, unpublished data, do not distribute
mRNA Signatures + DNA Copy Number + DNA mutations

Fan and Perou 2017, unpublished data, do not distribute
Clinical Data only

<table>
<thead>
<tr>
<th>Variable</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

Overall Survival (Probability)

Train

low 16/109
high 45/109

P = 2.12e-05

Test

low 11/69
high 27/69

P = 0.0308

Fan and Perou 2017, unpublished data, do not distribute
mRNA Signatures + DNA Copy Number + DNA mutations + Clinical Data

![Graphs showing overall survival probability](image)

**Train**
- C-index = 0.755
- low 12/109, high 49/109
  - P = 9.29e-07

**Test**
- C-index = 0.712
- low 10/69, high 28/69
  - P = 0.00074

**Test**
- C-index = 0.676
- low 14/69, high 24/69
  - P = 0.00946

**Clinical Data Only**
- C-index = 0.643
- low 11/69, high 27/69
  - P = 0.0308

---

Fan and Perou 2017, unpublished data, do not distribute
Systems Biology

Fan and Perou 2017, 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.
Homework

2/20/17 – Chuck Perou (Department of Genetics) – Introduction to Genomics and Big Data, and Cancer Subtype Class Discovery using gene expression data
- Katie Hoadley (Department of Genetics) – Introduction to TCGA Data Portal

Reading list
Eisen et al., PNAS 1998 (PMID:9843981)
Perou et al., NATURE 2000 (PMID:10963602)
Parker et al., JCO 2009 (PMID:19204204)

TCGA Breast Cancer Genomic Data Sites
https://tcga-data.nci.nih.gov/docs/publications/brca_2015/ (all open access TCGA Breast Cancer Data)
https://lbg.unc.edu/~hoadley/BRCA.817.rsemg.uqnorm.counts.txt (all 20,000 gene expression values)

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
Homework

Reading List
- Elsen et al., PNAS 1995 (PMID:5681231)
- Perou et al., NATURG 2001 (PMID:1192904)
- Parker et al., JCO 2005 (PMID:1626626)

TCGA Breast Cancer Sequencing Data
- http://www.participantcentral.org/tcgawww/proj/bcr/seq_dwnld.html
- Take the "class assignment" data, and make a hierarchical cluster.

1. Recommend using "Cluster 3.0" and under "adjust data" select log-transform and median center the data.
2. Add the "dendrogram" tab and select cluster (both genes and arrays). Similarity metric "centered", and cluster method "between-clus".

The data can then be merged and explored using "Gene Tree View". Look for and find the "proliferation-clusters", which were shown earlier in this presentation.

Gene Cluster 3.0

File:
Job name: EPGA.817,seqs,seqs,counts,filtered-CLUST
Data set has:
- 1879 Rows
- 817 Columns

Adjust Data:
- Log transform data
- Center genes
  - Mean
  - Median
- Normalize genes

Order of Operations:
- Log Transform
- Center Genes
- Normalize Genes
- Center Arrays
- Normalize Arrays

Done loading data.
Homework

Reading list:
- Rian et al., PNAS 1998 (PMID: 9543516)
- Perou et al., NATURE 2000 (PMID: 10953302)
- Parker et al., JCO 2000 (PMID: 1054504)

TGCA Breast Cancer Genetic Data Sites:
- https://www.sanger.ac.uk/sequencing/tgcabcr/
- https://www.sanger.ac.uk/sequencing/tgca/

Take the 'classification list' data, and make a hierarchical cluster:
1. Remove outliers using 'Cluster 1.0' and order 'adjust data' to select log transformation and median center the genes.
2. Next select the 'hierarchical' tab and select cluster (both genes and arrays), similarity metric 'correlation' and cluster method 'control linkage'.
3. The data can then be viewed and explored using 'Gene Therian'. Look for and find the 'proliferation cluster', which was showed earlier in this presentation.