Cancer Genomics and Class Discovery

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

2/27/17 – Katie Hoadley (Department of Genetics) – Multi-platform Data Analysis and Across Technology Data Integration

3/6/17 – Steve Marron (Department of Statistics and Operations Research) – Methods for Addressing Data Heterogeneity and Integration

3/13/17 – No Class

3/20/17 – Andrew Nobel (Department of Statistics and Operations Research) – Exploratory Analysis of Genomic Data

3/27/17 – Joel Parker (Department of Genetics) – Methods and Challenges in the Analysis of NextGen Sequence Data for DNAseq and RNAseq

4/3/17 – No Class

4/10/17 – In Class Student Presentations (70%) and 2-3 page Written Report (30%) covering a unique analysis performed on TCGA Cancer Genomics Data
Cancer Genomics and Class Discovery

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://lbg.unc.edu/~hoadley/BRCA.817.rsemg.uqnorm.counts.txt
https://lbg.unc.edu/~hoadley/BRCA.817.rsemg.uqnorm.counts.intrinsic.txt
What is Big Data?

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

microRNAs
1500/sample

Proteins
200/sample

DNA Copy Number
1,000,000/sample

DNA Methylation
450,000/sample

Somatic Mutations
25,000/sample
What is Big Data?

Gene Expression (mRNAs)

microRNAs

Proteins

DNA Copy Number

DNA Methylation

Somatic Mutations

~1.5 Million Data Points per sample X 10,000 samples = 15 Billion Data Points Total
TCGA Data is
1. 10,000 individual tumors
2. 33 diverse tumor types
3. Clinical and Pathology data
4. Molecular assays performed:
   1. DNA exomes (mutations)
   2. mRNA-seq (gene expression)
   3. microRNA-seq (microRNAs)
   4. DNA methylation arrays
   5. AFFY SNP arrays (genotypes and DNA copy number)
   6. RPPA protein data on ~60%
   7. H&E images of each tumor
<table>
<thead>
<tr>
<th>Rank</th>
<th>Cause of Death</th>
<th>Percent of Total Deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Causes</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>Heart Disease</td>
<td>23.5</td>
</tr>
<tr>
<td>2</td>
<td>Cancer</td>
<td>22.5</td>
</tr>
<tr>
<td>3</td>
<td>CLRD (Lung diseases)</td>
<td>5.7</td>
</tr>
<tr>
<td>4</td>
<td>Accidents</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>Cerebrovascular disease (stroke)</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>Alzheimer's disease</td>
<td>3.3</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes</td>
<td>2.9</td>
</tr>
<tr>
<td>8</td>
<td>Influenza and pneumonia</td>
<td>2.2</td>
</tr>
<tr>
<td>9</td>
<td>Kidney disease</td>
<td>1.8</td>
</tr>
<tr>
<td>10</td>
<td>Suicide</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>All other causes</td>
<td>26.4</td>
</tr>
</tbody>
</table>
### 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>

*Breast* | 232,340   | 29%

### 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>

*Breast* | 39,620    | 14%

*Note: The estimated new cases and deaths are based on the latest available data and are subject to change.*
Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative
T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),
All received neoadjuvant chemotherapy and adjuvant tamoxifen

Died after 9 months

Died after 17 months

Alive for 96 months
Central Dogma: DNA $\rightarrow$ RNA $\rightarrow$ Protein

- **DNA**
  - CCTGAGCCAACTATTGATGAA

- **RNA**
  - CCUGAGCCAAACUUUGAUGAA

- **Protein**
  - PEPTIDE
Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein


= high
= average
= low
Unsupervised cluster analysis of 10,000 grossly dissected human tumors
Proliferation Gene Set

Whitfield et al.
Common Markers of Proliferation
Nature Reviews Cancer
Vol. 6; 99-106 (2006)
PMID:12058064

- BUB1
- PLK1
- Thymidylate Synthetase
- EZH2
- DNA Polymerase alpha
- Cyclin A2, B2, E1
- Tubulin
- BRCA1, 2
- MCM2, 3, 5, 6, 7, 8, 10
- Forkhead Box M1
- MAD2
- Dihydrofolate reductase
- MYBL2
- Ki-67
- PTTG1
- Replication Factor C
- CENPA, E, F, H
- TOP2a
- STK6/15
- RAD51
- FANCA
- BRCA1, 2
- Ribonucleotide reductase
- CHEK1
- PCNA
- MSH2
- CDC1, 2, 7, 8, 20, 25
Chemotherapeutic drug targets include taxanes, anthracyclines, cytoxan, cis/carboplatin, capecitabene, gemcitabene, methotrexate, and more. 

- BUB1
- PLK1
- Thymidylate Synthetase = 5FU
- EZH2
- DNA Polymerase alpha
- Cyclin A2, B2, E1
- Tubulin = taxanes
- BRCA1, 2
- MCM2, 3, 5, 6, 7, 8, 10
- Forkhead Box M1
- MAD2
- Dihydrofolate reductase = methotrexate
- MYBL2
- Ki-67
- PTTG1
- Replication Factor C
- CENPA, E, F, H
- TOP2a = doxorubicin, etoposide
- STK6/15
- RAD51
- FANCA
- PCNA
- MSH2
- MAD2
- PTTG1
- Replication Factor C
- CENPA, E, F, H
- TOP2a = doxorubicin, etoposide
- STK6/15
- RAD51
- FANCA
- PCNA
- MSH2
- MAD2
- PTTG1
- Replication Factor C
- CENPA, E, F, H
- TOP2a = doxorubicin, etoposide
- STK6/15
- RAD51
- FANCA
- PCNA
- MSH2
9906 tumors mRNA-sequenced and uploaded to the CGHub public repository (https://cghub.ucsc.edu/), which represents >100 trillion bases. 33 tumor types studied including breast, bladder, colon, rectum, head & neck, renal clear cell & chromophobe, lung squamous & adenocarcinoma, melanoma, ovarian, glioblastoma, prostate, endometrial, thyroid, and many others. 10,000 tumors X 5000 genes.
A Complex Picture of Microarray/Gene Expression Data Analysis

**Unsupervised Learning**
- Consensus Cluster
- Hierarchical Clustering
- Bi-clustering

**Supervised Learning**
- DWD
- SVM
- Centroids
- Elastic Net

**Missing Data Imputation**
- K-NN
- Row Mean
- Upper quantile
- RMA
- LOWESS

**Data Normalization**
- Image Analysis (microarray)
- Read Mapping (sequencing)

**Batch Correction**
- Hypothesis Testing
  - Bonferroni
  - FDR estimations
  - permutation

**Katie Hoadley**

**Andrew Nobel**

**Steve Marron**

**Joel Parker**

**Chuck Perou**
Advice for Interpreting Genomic Analysis Results (or Big Data Results in General)

1. Does the selected feature set make biological sense and/or have a common annotation theme?

2. What was the False-Discovery Rate?

3. Was there a training set and an independent test set?

4. Has the new feature(s) been tested in a multivariable analysis with previously known important features?

Note = All these questions are true for ANY multi-analyte assay/result, and also true for ANY single analyte assay/result
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
Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative
T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),

Died after 9 months

Died after 17 months

Alive for 96 months
Agendia MammaPrint Assay
(Agilent DNA Microarray)

Sporadic breast tumours
patients <55 years
tumour size <5 cm
lymph node negative (LN0)

Prognosis Reporter Genes -70 total

Distant metastases
<5 years (2.5)
Bad prognosis

No distant metastases
>5 years (8.7)
Good Prognosis

training set of
78 tumors

virgin set of
19 tumors

A prospective randomized study comparing the 70-gene signature with the common clinical-pathological criteria in selecting patients for adjuvant chemotherapy in breast cancer with 0-3 positive nodes.

**EORTC Trial 10041 (BIG 3-04), MIND-ACT**

Total accrual (updated on 10/14/2009):
enrollment completed = ~6000 patients


**RANDOMIZE**

- Clinical-pathological and 70-gene both **HIGH** risk
- Discordant cases
  - Clin-Path **HIGH**
  - 70-gene **LOW**
  - Clin-Path **LOW**
  - 70-gene **HIGH**

**Clinical-pathological and 70-gene both LOW risk**

**Use Clin-Path risk to decide Chemo or not**

**Use 70-gene risk to decide Chemo or not**

- Chemotherapy
- Endocrine therapy

Potential CT sparing in 10-15% pts
EORTC Trial 10041 (BIG 3-04), MIND- ACT

Total accrual (updated on 10/14/2009):
enrollment completed = ~6000 patients
http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp

A prospective randomized study comparing the 70-gene signature with the common clinical-pathological criteria in selecting patients for adjuvant chemotherapy in breast cancer with 0-3 positive nodes.

- Clinical-pathological and 70-gene both HIGH risk
- Discordant cases
  - Clin-Path HIGH
  - 70-gene LOW
  - Clin-Path LOW
  - 70-gene HIGH

RANDOMIZE
- Use Clin-Path risk to decide Chemo or not
- Use 70-gene risk to decide Chemo or not

Chemotherapy

- Clinical-pathological and 70-gene both LOW risk

- Potential CT sparing in 10-15% pts

Endocrine therapy

Figure 3. Survival without Distant Metastasis in the Four Risk Groups.
The analysis includes all enrolled patients, and the risk groups are based on corrected risk. The time-to-event curves were estimated by means of the Kaplan–Meier method.
Genomic Microarray-based platforms (and NGS) can currently assay for

1. Gene Expression: typically thought of as mRNA/protein coding genes = ~25,000 genes

2. MicroRNAs: important class of small regulatory RNA molecules = ~1500

3. Long Non-Coding RNAs (lncRNAs): are mRNAs with PolyA tails and exons, but do not code for proteins = 40,000 - 50,000

4. DNA Methylation = 450,000 methylation sites

5. Germline DNA Polymorphisms (>1,000,000 SNPs) and Tumor DNA Copy Number Changes
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
Quantitative RT-PCR

Keratin 5

Fluorescence (530nm)

Cycle Number

100ng
10ng
1ng
0.1ng
0.01ng

ABI 7900HT Fast Real-Time PCR System

Roche Lightcycler 480
Formalin-Fixed Paraffin Embedded (FFPE) RNAs pose a unique challenge for nucleic acid based studies.
A Multi-gene Assay to Predict Recurrence of Tamoxifen-Treated Node-Negative Breast Cancer.

Figure 1. Panel of 21 Genes and the Recurrence-Score Algorithm.
The recurrence score on a scale from 0 to 100 is derived from the reference- normalized expression measurements in four steps. First, expression for each gene is normalized relative to the expression of the five reference genes (ACTB [the gene encoding β-actin], GAPDH, GUS, RPLPO, and TFRC). Reference- normalized expression measurements range from 0 to 15, with a 1-unit increase reflecting approximately a doubling of RNA. Genes are grouped on the basis of function, correlated expression, or both. Second, the GRB7, ER, proliferation, and invasion group scores are calculated from individual gene-expression measurements, as follows: GRB7 group score = 0.9 x GRB7 + 0.1 x HER2 (if the result is less than 8, then the GRB7 group score is considered 8); ER group score = (0.8 x ER + 1.2 x PGR + BCL2 + SCUBE2) / 4; proliferation group score = Survivin + K67 + MYBL2 + CCNB1 [the gene encoding cyclin B1] + STK15 + 5 (if the result is less than 6.5, then the proliferation group score is considered 6.5); and invasion group score = (CTSL2 [the gene encoding cathepsin L2] + MMP11 [the gene encoding stromelysin 3]) / 2. The unscaled recurrence score (RS_U) is calculated with the use of coefficients that are predefined on the basis of regression analysis of gene expression and recurrence in the three training studies*+**: RS_U = 0.47 x GRB7 group score - 0.34 x ER group score + 1.04 x proliferation group score + 0.10 x invasion group score + 0.05 x CD68 - 0.08 x GSTM1 - 0.07 x BAG1. A plus sign indicates that increased expression is associated with an increased risk of recurrence, and a minus sign indicates that increased expression is associated with a decreased risk of recurrence. Fourth, the recurrence score (RS) is rescaled from the unscaled recurrence score, as follows: RS = 0 if RS_U < 0; RS = 20 x (RS_U - 6.7) if 0 ≤ RS_U ≤ 100; and RS = 100 if RS_U > 100.

Figure 2. Likelihood of Distant Recurrence, According to Recurrence-Score Categories.
A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher. There were 28 recurrences in the low-risk group, 25 in the intermediate-risk group, and 56 in the high-risk group. The difference among the groups is significant (P<0.001).
Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor-Positive Breast Cancer
NSABP B-20 (651 tumor FFPE samples)

**Fig 2.** Kaplan-Meier plots for distant recurrence comparing treatment with tamoxifen (Tam) alone versus treatment with tamoxifen plus chemotherapy (Tam + chemo). (A) All patients; (B) low risk (recurrence score [RS] < 18); (C) intermediate risk (RS 18-30); (D) high risk (RS ≥ 31). The number of patients at risk and the number of distant recurrences (in parentheses) are provided below each part of the figure.
TAILORx Study Design

ECOG/Inter-group
PI: J. A. Sparano

Accrual completed on Oct 25th 2010, 10,000 total. expected to report in 2018?

Pre-REGISTER
ONCOTYPE DX ASSAY
REGISTER Specimen Banking

Secondary Study Group 1
RS < 11
~29% of Population

ARM A Hormonal Therapy Alone

Primary Study Group
RS 11-25
~44% of Population

RANDOMIZE
Stratification Factors:
Tumor Size, Menopausal Status,
Planned Chemo, Planned Radiation

Secondary Study Group 2
RS > 25
~27% of Population

ARM D Chemotherapy Plus Hormonal Therapy

ARM B Hormonal Therapy Alone

ARM C Chemotherapy Plus Hormonal Therapy
Intrinsic Subtypes of Breast Cancer

- Normal Breast
- Claudin-low
- HER2-enriched
- Luminal A
- Luminal B
- Basal-like

Intriguing visualization with data scatter plots and survival curves comparing different breast cancer subtypes.
Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),
All received neoadjuvant chemotherapy and adjuvant tamoxifen

Died after 9 months: LumB

Died after 17 months: Basal-like

Alive for 96 months: LumA
1. The qRT-PCR assay consists of 50 genes and 5 centroids (provided at https://genome.unc.edu/)

2. The Cross Validation (CV) classification concordance of the 50 genes by qRT-PCR compared with 2000 genes by microarray was 93%

3. The assay works using RNA from FFPE materials or fresh frozen tissues
Step 1: cluster analysis using ~2000 gene intrinsic list, which was three previous intrinsic lists combined.

Step 2: use “SigClust” to objectively identify significant groups/subtypes, and the prototypical tumors of each group/subtype.

Step 3: use cross validation classification concordance testing to shrink the gene list from ~2000 genes to 50 genes, using “ClaNC” algorithm to balance gene selection (equal # of genes per class).

Step 4: use Nearest Centroid Predictor with 5 subtypes and 50 genes (PAM50), by qRT-PCR with FFPE derived RNA.
710 node negative breast cancer patients who received no systemic adjuvant therapy as a test set (pure prognosis)

Table 2. Models of Relapse Free Survival (untreated)

<table>
<thead>
<tr>
<th>Model</th>
<th>Variable</th>
<th>A</th>
<th>Hazard Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Basal-like*</td>
<td></td>
<td>1.33</td>
<td>0.330</td>
</tr>
<tr>
<td></td>
<td>HER2-enriched*</td>
<td></td>
<td>2.53</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>Luminal B*</td>
<td></td>
<td>2.43</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>ER Status~</td>
<td></td>
<td>0.83</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Tumor Size†</td>
<td></td>
<td>1.36</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>Node Status‡</td>
<td></td>
<td>1.75</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>Histologic Grade^</td>
<td></td>
<td>1.40</td>
<td>0.0042</td>
</tr>
<tr>
<td></td>
<td>Full vs Subtype≈</td>
<td></td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Full vs Clinical¥</td>
<td></td>
<td>&lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>Hazard Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1.79</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.25</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.88</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.83</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.43</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.72</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>Hazard Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1.58</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.90</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.54</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.83</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.57</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Parker et al. JCO 2009 (PMID:19204204)
Similarity to the subtypes are used as variables in the prognostic model where the outcome is Risk of Recurrence (ROR):

(Model 1) \[ \text{ROR-S} = \beta_1 \cdot \text{Basal} + \beta_2 \cdot \text{HER2} + \beta_3 \cdot \text{LumA} + \beta_4 \cdot \text{LumB} \]

(Model 2) \[ \text{ROR-T} = \beta_1 \cdot \text{Basal} + \beta_2 \cdot \text{HER2} + \beta_3 \cdot \text{LumA} + \beta_4 \cdot \text{LumB} + \beta_5 \cdot \text{Size} \]

(Model 3) \[ \text{ROR-PT} = \beta_1 \cdot \text{Basal} + \beta_2 \cdot \text{HER2} + \beta_3 \cdot \text{LumA} + \beta_4 \cdot \text{LumB} + \beta_5 \cdot \text{Size} + \beta_6 \cdot \text{Proliferation} \]

Weights (\( \beta \)) for each variable are learned from a training data set using a Cox proportional hazards model with Ridge Regression\(^1\)

The weighted sum is assigned as the ROR score for a test case and a threshold may be applied for class assignment.

\(^1\)Ridge regression with Cox model: Tibshirani, Statistics in Medicine, 1997 (PMID:9044528) and Bovelstad et al. Bioinformatics 2007 (PMID:17553857)
Prognostic Risk Classification Strategy (ROR)

N=558 no adjuvant systemic therapy and node negative test cases

ROR-T thresholds determined from training cases

Parker et al. JCO 2009 (PMID:19204204)
The c-index is the proportion of all pairs of subjects whose survival time can be ordered such that the subject with the higher predicted survival is the one who survived longer” (taken from Harrell, Regression Modeling Strategies, Springer Series in Statistics).
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
Scientific Reasons for PAM50 platform change to NanoString nCounter (~2010)

- Easier to use than PCR or microarrays
- Fast and simple workflow
- Compatible with a distributed test and prefabricated kits (IVD)
- Compatible with RNA coming from FFPE materials
Schematic of gene detection probe system

1. Capture and Reporter probes are mixed with RNAs for hybridization
2. Gene specific oligos are covalently attached by ligation reaction
3. RNA segments are annealed to surface via 3' repeats and imaged
4. 1-800 genes can be assayed per sample