Machine learning methods for accurate prediction of cancer metastasis

Jianhua Ruan, PhD

Dept of Computer Science, Univ of Texas at San Antonio
Cancer Therapy and Research Center, UTHSCSA
Cancer metastasis

- Metastasis is the spread of cancer from one organ (primary) to another (secondary)
- Majority of cancer-related death due to metastasis
- Primary cancer is usually removed by surgery
- When metastasis is likely, adjuvant therapy including radiation therapy and chemotherapy is often prescribed
  - Severe side affect
- Challenge: how to predict metastasis at an earlier stage of cancer?
Prediction of metastatic breast cancer

- Clinical and pathological risk factors including
  - Tumor size
  - Patient age
  - Steroid receptor status

- Traditional risk factor are not strongly predictive for intermediate risk patient group

- Approximately 70-80% of lymph node-negative cancer patients undergo unnecessary adjuvant chemotherapy
Breast Cancer Metastasis

Degree of Malig

< 3

Tumor Size

< 15

Age

no meta 4
meta 1

no meta 125
meta 39

no meta 32
meta 0

>= 15

Involved Nodes

< 3

no meta 30
rmeta 18

meta 27
no meta 10

>= 3

Age

no meta 4
meta 1

no meta 32
meta 0
Microarray based prediction of metastasis

- Microarray: measures gene expression (activity) level
- Predictive models trained using data from patients of known outcomes
  - Genes as predictor variables
  - Disease outcome as response variable
  - (Optional) statistical test such as t-test applied to select predictor variables
  - Predictive models include logistic regression, decision tree, support vector machines, neural networks, etc.
    - E.g. \( p = 0.03g_1 + 0.01g_2 - 0.02g_3 - 0.01g_4 \ldots \)
- Commercialized test kits available
  - Mammaprint
  - Oncotype DX
Class A (e.g. non-metastatic tumor)

1  2  3  4  5

Pattern for class A

Class B (e.g. metastatic tumor)

6  7  8  9  10

Pattern for class B

? more like pattern A or like pattern B?

Source: “Practical Microarray Analysis”, Presentation by Benedikt Brors, German Cancer Research Center

Trinity Biomath
Centroid-based classifier

- **Model Training:** Based on the training data calculate the centroid for each class.

- **Classification:**
  1. Given a data point, calculate the distance between the point and each of the class centroids.
  2. Assign the point to the closest class.
K-Nearest-Neighbour classifier

- Model Training: none
- Classification:
  - Given a data point, locate K nearest points.
  - Returns the most common class label among the k points nearest to x
- We usually set $K > 1$ to avoid outliers
- Variations:
  - Can also use a radius threshold rather than K.
  - We can also set a weight for each neighbour that takes into account how far it is from the query point
Performance evaluation

- How predictive is the model we learned?
- Error on the training data is *not* a good indicator of performance on future data
  - *Q: Why?*
  - *A: Because new data will probably not be **exactly** the same as the training data!*
- Overfitting – fitting the training data too precisely - usually leads to poor results on new data
Evaluation on “LARGE” data

- If many (thousands) of examples are available, including several hundred examples from each class, then how can we evaluate our classifier method?
- A simple evaluation is sufficient
  - Randomly split data into training and test sets (usually 2/3 for train, 1/3 for test)
- Build a classifier using the train set and evaluate it using the test set.
Classification Step 1:
Split data into train and test sets

THE PAST
Results Known

Data → \[+\] → Training set

\[+\] → [ ] → Testing set
Classification Step 2: Build a model on a training set

THE PAST
Results Known

Data → + + + + + +

Training set → Model Builder

Testing set
Classification Step 3: Evaluate on test set

Results Known

Training set

Model Builder

Evaluate

Predictions
A note on parameter tuning

- It is important that the test data is not used in any way to create the classifier.
- Some learning schemes operate in two stages:
  - Stage 1: builds the basic structure
  - Stage 2: optimizes parameter settings
- The test data can’t be used for parameter tuning!
- Proper procedure uses three sets: training data, validation data, and test data
  - Validation data is used to optimize parameters
Classification: Train, Validation, Test split

Model Builder

Data

Results Known

Model Builder

Validation set

Training set

Final Test Set

Final Model

Final Evaluation
Cross-validation

- Cross-validation more useful in small datasets
  - First step: data is split into $k$ subsets of equal size
  - Second step: each subset in turn is used for testing and the remainder for training
- This is called $k$-fold cross-validation
- Often the subsets are stratified before the cross-validation is performed
- The error estimates are averaged to yield an overall error estimate
Cross-validation example:

- Break up data into groups of the same size

- Hold aside one group for testing and use the rest to build model

- Repeat
Limitations of current microarray-based prediction of metastasis

- Still low accuracy
- Unstable genes / models among different studies
  - ~70 markers, 60-70% accurate to predict metastasis
  - Only 3 of those markers are in common
- Possible causes
  - Noisy data
  - Downstream changes vs true causal factors
  - Heterogeneous experimental platforms
  - Heterogeneous cancer subtypes
  - Small sample size vs large number of genes
Limitations of current microarray-based prediction of MBC

- Still low accuracy
- Unstable genes / models among different studies
  - ~70 markers, 60-70% accurate to predict metastasis
  - Only 3 of those markers are in common
- Possible causes
  - Noisy data
  - Downstream changes vs true causal factors
  - Heterogeneous experimental platforms
  - Heterogeneous cancer subtypes
    - => Personized model training and selection
    - Small sample size vs large number of genes;
    => Consider gene-gene networks
Breast cancer subtypes

- Peru and colleagues defined five subtypes for breast cancer:
  - Luminal A, Luminal B, ErbB2, Basal, Normal
  - Recently found another subtype known as Claudia-low

- Other studies confirm Basal and ErbB2 but other subtypes not confirmed or consistent

- Recent studies suggest that molecular subtype of cancer may be a continuum rather than discrete
Subtype-specific models

- Different models for different subtypes
  - Overall slightly worse accuracy
  - Improvement for Luminal B and ErbB2 subtypes
  - Significantly poorer for normal and basal

- Possible causes:
  - Subtype definition is ambiguous
  - Sub-subtypes?
  - Fewer training samples for each subtype
    => risk of overfitting
Personalized cancer prognosis

- Personalized model construction
  - Build an ensemble of models trained on selected subsets of patients similar to some target patients
  - (Train doctors specialized for different patient characteristics)

- Personalized model selection
  - Find a subset of models best for an individual
  - (Doctor-patient matching)

- Committee-based decision making
  - e.g. confidence-weighted voting
Personalized model construction

Training data selection
Each patient (P) is considered as a different subtype
Implicitly defined by a group of similar patients, which are selected using random walk on the patient-patient network
Random walk based neighbor selection

\[ P = (1-c) \times A \times P + c \times P_0 \]

- Takes into account distance and topology
- Automatically identify neighbors in the same “cluster”
- Popular and validated method in machine learning
Random walk based neighbor selection

Different cutoffs lead to different number of selected neighbors
Random walk based neighbor selection

Different cutoffs lead to different number of selected neighbors
Random walk based neighbor selection

Different cutoffs lead to different number of selected neighbors
Personalized model construction

- Multiple models trained for each patient $P$ using selected neighbors
- Incorrect/inaccurate models removed

Jahid et al., Bioinformatics 2014
Personalized model selection and decision making

- Identify **past patients** (with known outcomes) that have similar molecular characteristics as the testing patient
- Reuse the models that worked on these patients
- Final prediction made via weighted voting

Jahid et al., Bioinformatics 2014
Cross-validation on NKI Dataset

- Improvement for all subtypes
  - Most significant improvement for Basal and Normal – hardest cases
- For luminal subtype, known to be the least lethal one, at 75% sensitivity
  - 31.5% FPR for PC-classifier vs 43.1% FPR for standard SVM
  - 11.6% of luminal patients can avoid unnecessary adjuvant chemotherapy
Cross-dataset Performance

- Performance on Wang and UNC datasets using models trained from NKI dataset

Jahid et al., Bioinformatics 2014
Gene weights in different models

Gene clustering

Patient clustering

Metastasis status

Known subtype
Significance of Top-Ranked Genes

- Identify top genes in different clusters
- Identify their association with cancer using literature mining
  - Search keyword in PubMed abstracts
- Many top genes already known to be metastasis related
  - MMP9, PDGFRA, CCL21 etc
Significance of Top-Ranked Genes

- **Observation:**
  - Many genes show subtype specificity and cannot be identified by standard SVM classifier
  - ID1, HEY1, MST1R have high rank in Basal group but low rank in standard SVM classifier
  - ID1: well known mediator of breast cancer lung metastatic patient for Basal group
  - HEY1: target gene for Notch signaling inhibitor for basal group

<table>
<thead>
<tr>
<th>Gene name</th>
<th>PubMed hits</th>
<th>Standard SVM</th>
<th>Basal</th>
<th>LumB/Erbb2</th>
<th>LumA/LumB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>38</td>
<td>878</td>
<td>19</td>
<td>770</td>
<td>682</td>
</tr>
<tr>
<td>ATP1B1</td>
<td>5</td>
<td>88</td>
<td>22</td>
<td>459</td>
<td>1112</td>
</tr>
<tr>
<td>CCL19</td>
<td>33</td>
<td>252</td>
<td>10</td>
<td>704</td>
<td>692</td>
</tr>
<tr>
<td>HEY1</td>
<td>4</td>
<td>372</td>
<td>8</td>
<td>368</td>
<td>680</td>
</tr>
<tr>
<td>MST1R</td>
<td>3</td>
<td>783</td>
<td>6</td>
<td>741</td>
<td>1456</td>
</tr>
<tr>
<td>ANO1</td>
<td>3</td>
<td>490</td>
<td>9</td>
<td>528</td>
<td>652</td>
</tr>
<tr>
<td>FABP7</td>
<td>3</td>
<td>358</td>
<td>10</td>
<td>371</td>
<td>380</td>
</tr>
<tr>
<td>KLK4</td>
<td>6</td>
<td>900</td>
<td>6</td>
<td>916</td>
<td>144</td>
</tr>
<tr>
<td>KLK6</td>
<td>9</td>
<td>645</td>
<td>29</td>
<td>229</td>
<td>759</td>
</tr>
<tr>
<td>NDGR1</td>
<td>56</td>
<td>414</td>
<td>1383</td>
<td>22</td>
<td>33</td>
</tr>
<tr>
<td>FLT3</td>
<td>48</td>
<td>395</td>
<td>873</td>
<td>199</td>
<td>26</td>
</tr>
<tr>
<td>SPP1</td>
<td>26</td>
<td>324</td>
<td>334</td>
<td>47</td>
<td>18</td>
</tr>
<tr>
<td>TFF1</td>
<td>24</td>
<td>99</td>
<td>181</td>
<td>30</td>
<td>231</td>
</tr>
<tr>
<td>HOXB13</td>
<td>7</td>
<td>89</td>
<td>659</td>
<td>21</td>
<td>1214</td>
</tr>
<tr>
<td>ITGA2</td>
<td>5</td>
<td>64</td>
<td>247</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>PEG10</td>
<td>5</td>
<td>97</td>
<td>748</td>
<td>28</td>
<td>53</td>
</tr>
<tr>
<td>CKB</td>
<td>5</td>
<td>286</td>
<td>16</td>
<td>27</td>
<td>1254</td>
</tr>
<tr>
<td>IGFBP3</td>
<td>2</td>
<td>277</td>
<td>25</td>
<td>545</td>
<td>29</td>
</tr>
<tr>
<td>PDGFRA</td>
<td>144</td>
<td>43</td>
<td>26</td>
<td>82</td>
<td>1131</td>
</tr>
<tr>
<td>GP5</td>
<td>12</td>
<td>23</td>
<td>18</td>
<td>120</td>
<td>311</td>
</tr>
<tr>
<td>PIBX1</td>
<td>8</td>
<td>28</td>
<td>407</td>
<td>134</td>
<td>21</td>
</tr>
<tr>
<td>NUPR1</td>
<td>6</td>
<td>15</td>
<td>348</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>ANXA3</td>
<td>4</td>
<td>18</td>
<td>1519</td>
<td>51</td>
<td>3</td>
</tr>
<tr>
<td>BMP1B</td>
<td>2</td>
<td>30</td>
<td>1086</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>MMP9</td>
<td>323</td>
<td>7</td>
<td>572</td>
<td>29</td>
<td>4</td>
</tr>
<tr>
<td>ADAM8</td>
<td>15</td>
<td>24</td>
<td>160</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>CYP1B1</td>
<td>14</td>
<td>26</td>
<td>43</td>
<td>34</td>
<td>170</td>
</tr>
<tr>
<td>IGFBP5</td>
<td>11</td>
<td>33</td>
<td>23</td>
<td>11</td>
<td>104</td>
</tr>
<tr>
<td>PRAME</td>
<td>9</td>
<td>35</td>
<td>20</td>
<td>4</td>
<td>409</td>
</tr>
<tr>
<td>CCL21</td>
<td>58</td>
<td>20</td>
<td>227</td>
<td>519</td>
<td>79</td>
</tr>
</tbody>
</table>
Significance of Top-Ranked Genes

- **Observation:**
  - **NDGR1:** Known to be related to Luminal subtype
  - **TFF1:** Known to be related to luminal stability
  - **PDGFRA:** Drug-target for basal like tumor
  - **BMPR1B:** Associated with ER-positive breast cancer subtype
Personalized cancer prognosis - Summary

- Personalized models provide better classification accuracy for both in-data and cross-data evaluation.
- Clustering of personalized models offer biological insights for progression of different cancer subtypes.

Limitations of current microarray-based prediction of MBC

- Still low accuracy
- Unstable genes / models among different studies
  - ~70 markers, 60-70% accurate to predict metastasis
  - Only 3 of those markers are in common

- Possible causes
  - Noisy data
  - Downstream changes vs true causal factors
  - Heterogeneous experimental platforms
  - Heterogeneous cancer subtypes
  - Personized model training and selection
  - Small sample size vs large number of genes
  - Consider gene-gene networks
Why networks for cancer?

- Candidate cancer genes
  - Too many to understand the biology
  - Driver vs. passenger

- Gene-centric classifier
  - Not robust
  - Lacks biological insight

- Network types
  - Protein-protein
  - Protein-DNA
  - miRNA-mRNA
  - Signaling networks
  - ...
Connectors as additional biomarkers

- Many cancer genes are in close proximity to disrupted subnetworks
- Goal: given list of genes disrupted in cancer, find a small subnetwork that connects them

Steiner tree problem
Data

- 60 Endometrial (womb) cancer patients
  - 16 recurrent (R) in 3 years
  - 44 non-recurrent (NR)
  - 12 normal control
- Global methylation pattern surveyed by MBDcap-seq at OSU/UTHSCSA
- 4214 CpG islands differentially methylated between cancer and control
- Among them, 135 genes differentially methylated (DM) between R and NR
Diffrentially methylated (DM) gene
Epigenetic connector (EC) gene
Protein-protein interaction
Methylation of DM and EC genes

Connectors include both weak DM genes in the same pathway as the DM genes, and genes that are not differentially methylated but facilitate cross-talks between DM pathways.
<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>PValue</th>
<th>List Total</th>
<th>Pop Hits</th>
<th>Pop Total</th>
<th>Fold Enrichment</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa05200:Pathways in cancer</td>
<td>70</td>
<td>1.24E-24</td>
<td>278</td>
<td>328</td>
<td>5085</td>
<td>3.90</td>
<td>1.4E-21</td>
</tr>
<tr>
<td>hsa04722:Neurotrophin signaling pathway</td>
<td>36</td>
<td>7.48E-17</td>
<td>278</td>
<td>124</td>
<td>5085</td>
<td>5.31</td>
<td>1.2E-13</td>
</tr>
<tr>
<td>hsa04350:TGF-beta signaling pathway</td>
<td>29</td>
<td>3.02E-15</td>
<td>278</td>
<td>87</td>
<td>5085</td>
<td>6.10</td>
<td>3.4E-12</td>
</tr>
<tr>
<td>hsa04012:ErbB signaling pathway</td>
<td>26</td>
<td>1.80E-12</td>
<td>278</td>
<td>87</td>
<td>5085</td>
<td>5.47</td>
<td>2.1E-09</td>
</tr>
<tr>
<td>hsa04510:Focal adhesion</td>
<td>39</td>
<td>4.69E-12</td>
<td>278</td>
<td>201</td>
<td>5085</td>
<td>3.55</td>
<td>5.4E-09</td>
</tr>
<tr>
<td>hsa04540:Gap junction</td>
<td>23</td>
<td>1.02E-09</td>
<td>278</td>
<td>89</td>
<td>5085</td>
<td>4.73</td>
<td>1.2E-06</td>
</tr>
<tr>
<td>hsa04110:Cell cycle</td>
<td>27</td>
<td>1.68E-09</td>
<td>278</td>
<td>125</td>
<td>5085</td>
<td>3.95</td>
<td>1.9E-06</td>
</tr>
<tr>
<td>hsa04912:GnRH signaling pathway</td>
<td>23</td>
<td>7.20E-09</td>
<td>278</td>
<td>98</td>
<td>5085</td>
<td>4.29</td>
<td>8.2E-06</td>
</tr>
<tr>
<td>hsa04010:MAPK signaling pathway</td>
<td>40</td>
<td>7.31E-09</td>
<td>278</td>
<td>267</td>
<td>5085</td>
<td>2.74</td>
<td>8.4E-06</td>
</tr>
<tr>
<td>hsa04660:T cell receptor signaling pathway</td>
<td>24</td>
<td>9.54E-09</td>
<td>278</td>
<td>108</td>
<td>5085</td>
<td>4.06</td>
<td>1.1E-05</td>
</tr>
<tr>
<td>hsa04144:Endocytosis</td>
<td>32</td>
<td>9.81E-09</td>
<td>278</td>
<td>184</td>
<td>5085</td>
<td>3.18</td>
<td>1.1E-05</td>
</tr>
<tr>
<td>hsa04062:Chemokine signaling pathway</td>
<td>32</td>
<td>1.46E-08</td>
<td>278</td>
<td>187</td>
<td>5085</td>
<td>3.13</td>
<td>1.7E-05</td>
</tr>
<tr>
<td>hsa04520:Adherens junction</td>
<td>20</td>
<td>1.52E-08</td>
<td>278</td>
<td>77</td>
<td>5085</td>
<td>4.75</td>
<td>1.7E-05</td>
</tr>
<tr>
<td>hsa04664:Fc epsilon RI signaling pathway</td>
<td>20</td>
<td>1.92E-08</td>
<td>278</td>
<td>78</td>
<td>5085</td>
<td>4.69</td>
<td>2.2E-05</td>
</tr>
<tr>
<td>hsa04920:Adipocytokine signaling pathway</td>
<td>18</td>
<td>5.73E-08</td>
<td>278</td>
<td>67</td>
<td>5085</td>
<td>4.91</td>
<td>6.5E-05</td>
</tr>
<tr>
<td>hsa04620:Toll-like receptor signaling path</td>
<td>21</td>
<td>3.26E-07</td>
<td>278</td>
<td>101</td>
<td>5085</td>
<td>3.80</td>
<td>3.7E-04</td>
</tr>
<tr>
<td>hsa04662:B cell receptor signaling pathway</td>
<td>18</td>
<td>3.39E-07</td>
<td>278</td>
<td>75</td>
<td>5085</td>
<td>4.39</td>
<td>3.9E-04</td>
</tr>
<tr>
<td>hsa04370:VEGF signaling pathway</td>
<td>17</td>
<td>1.81E-06</td>
<td>278</td>
<td>75</td>
<td>5085</td>
<td>4.15</td>
<td>2.1E-03</td>
</tr>
<tr>
<td>hsa04020:Calcium signaling pathway</td>
<td>27</td>
<td>2.30E-06</td>
<td>278</td>
<td>176</td>
<td>5085</td>
<td>2.81</td>
<td>2.6E-03</td>
</tr>
<tr>
<td>hsa04530:Tight junction</td>
<td>23</td>
<td>2.43E-06</td>
<td>278</td>
<td>134</td>
<td>5085</td>
<td>3.14</td>
<td>2.8E-03</td>
</tr>
<tr>
<td>hsa05213:Endometrial cancer</td>
<td>14</td>
<td>2.62E-06</td>
<td>278</td>
<td>52</td>
<td>5085</td>
<td>4.92</td>
<td>3.0E-03</td>
</tr>
<tr>
<td>hsa04621:NOD-like receptor signaling path</td>
<td>15</td>
<td>3.98E-06</td>
<td>278</td>
<td>62</td>
<td>5085</td>
<td>4.43</td>
<td>4.6E-03</td>
</tr>
<tr>
<td>hsa04622:RIG-I-like receptor signaling path</td>
<td>15</td>
<td>2.11E-05</td>
<td>278</td>
<td>71</td>
<td>5085</td>
<td>3.86</td>
<td>2.4E-02</td>
</tr>
<tr>
<td>hsa04910:Insulin signaling pathway</td>
<td>20</td>
<td>1.09E-04</td>
<td>278</td>
<td>135</td>
<td>5085</td>
<td>2.71</td>
<td>1.3E-01</td>
</tr>
<tr>
<td>hsa04210:Apoptosis</td>
<td>15</td>
<td>2.16E-04</td>
<td>278</td>
<td>87</td>
<td>5085</td>
<td>3.15</td>
<td>2.5E-01</td>
</tr>
<tr>
<td>hsa04310:Wnt signaling pathway</td>
<td>20</td>
<td>4.81E-04</td>
<td>278</td>
<td>151</td>
<td>5085</td>
<td>2.42</td>
<td>5.5E-01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>PValue</th>
<th>List Total</th>
<th>Pop Hits</th>
<th>Pop Total</th>
<th>Fold Enrichment</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04080:Neuroactive ligand-receptor inter</td>
<td>11</td>
<td>2.12E-06</td>
<td>33</td>
<td>256</td>
<td>5085</td>
<td>6.62</td>
<td>2.2E-03</td>
</tr>
</tbody>
</table>
Network-based classification

- **Motivation**

Want: $A \approx B \neq C$

But: $A \cdot B = 5$

$A \cdot C = 5$

$B \cdot C = 4$
Strategy: information diffusion

\[ P = (1-c) \times A \times P + c \times P_0 \]
Utilizes subnetworks for classification

Network-based classification

Selected Genes

Outcomes

All patients with known outcomes
Methylation of EC genes

Before info diffusion
No genes passed statistical significance test between R and NR

After info diffusion
203 (43%) of genes passed statistical significance test (p < 0.02)
## Classification accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC</th>
<th>Kappa</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>0.646 ± 0.034</td>
<td>0.300 ± 0.071</td>
<td>0.734 ± 0.031</td>
<td>0.458 ± 0.056</td>
<td>0.834 ± 0.036</td>
</tr>
<tr>
<td>EC</td>
<td>0.464 ± 0.026</td>
<td>-0.089 ± 0.063</td>
<td>0.653 ± 0.028</td>
<td>0.059 ± 0.042</td>
<td>0.868 ± 0.037</td>
</tr>
<tr>
<td>EC*</td>
<td>0.629 ± 0.039</td>
<td>0.301 ± 0.087</td>
<td>0.767 ± 0.029</td>
<td>0.333 ± 0.076</td>
<td>0.924 ± 0.032</td>
</tr>
<tr>
<td>EC#</td>
<td>0.516 ± 0.037</td>
<td>0.035 ± 0.082</td>
<td>0.660 ± 0.033</td>
<td>0.207 ± 0.061</td>
<td>0.825 ± 0.036</td>
</tr>
<tr>
<td>DM+EC</td>
<td>0.700 ± 0.044</td>
<td>0.453 ± 0.096</td>
<td>0.812 ± 0.033</td>
<td>0.459 ± 0.075</td>
<td>0.940 ± 0.027</td>
</tr>
<tr>
<td>DM+</td>
<td>0.662 ± 0.038</td>
<td>0.358 ± 0.082</td>
<td>0.773 ± 0.029</td>
<td>0.424 ± 0.069</td>
<td>0.900 ± 0.030</td>
</tr>
<tr>
<td>DM+rand</td>
<td>0.658 ± 0.039</td>
<td>0.326 ± 0.079</td>
<td>0.744 ± 0.030</td>
<td>0.474 ± 0.070</td>
<td>0.840 ± 0.034</td>
</tr>
<tr>
<td>DM+EC*</td>
<td>0.731 ± 0.038</td>
<td>0.513 ± 0.078</td>
<td>0.829 ± 0.026</td>
<td>0.522 ± 0.070</td>
<td>0.941 ± 0.019</td>
</tr>
<tr>
<td>DM+EC#</td>
<td>0.698 ± 0.036</td>
<td>0.442 ± 0.078</td>
<td>0.805 ± 0.028</td>
<td>0.468 ± 0.063</td>
<td>0.927 ± 0.026</td>
</tr>
<tr>
<td>KEGG</td>
<td>0.548 ± 0.018</td>
<td>0.116 ± 0.046</td>
<td>0.713 ± 0.023</td>
<td>0.194 ± 0.020</td>
<td>0.902 ± 0.031</td>
</tr>
<tr>
<td>DS</td>
<td>0.486 ± 0.040</td>
<td>-0.028 ± 0.078</td>
<td>0.596 ± 0.038</td>
<td>0.249 ± 0.060</td>
<td>0.723 ± 0.045</td>
</tr>
<tr>
<td>DS*</td>
<td>0.656 ± 0.033</td>
<td>0.307 ± 0.063</td>
<td>0.725 ± 0.026</td>
<td>0.507 ± 0.059</td>
<td>0.805 ± 0.030</td>
</tr>
<tr>
<td>DM+DS</td>
<td>0.671 ± 0.042</td>
<td>0.369 ± 0.089</td>
<td>0.772 ± 0.032</td>
<td>0.457 ± 0.073</td>
<td>0.886 ± 0.030</td>
</tr>
<tr>
<td>DM+DS*</td>
<td>0.710 ± 0.035</td>
<td>0.426 ± 0.073</td>
<td>0.777 ± 0.030</td>
<td>0.568 ± 0.053</td>
<td>0.854 ± 0.032</td>
</tr>
</tbody>
</table>

- EC*: random walk on real PPI
- EC#: random walk on randomized PPI

Ruan et. al. Genomics 2016
Literature validation

Group III genes: ECs not differentially methylated, significance due to info diffusion

Table shows number of pubmed abstracts retrieved using gene name + “metastasis or metastatic”, or gene name + “epigenetic or methylation”

<table>
<thead>
<tr>
<th>Recurrence markers</th>
<th>Metastasis</th>
<th>Epigenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPHB2</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>COIL</td>
<td>209</td>
<td>65</td>
</tr>
<tr>
<td>STAU1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GSK3B</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>ID3</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>ID2</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>MCM6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BRCA1</td>
<td>297</td>
<td>356</td>
</tr>
<tr>
<td>SSTR2</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>SSTR3</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non-recurrence markers</th>
<th>Metastasis</th>
<th>Epigenetic</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARP1</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>TXNDC17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AVPR1A</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>SPHK1</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>TLE1</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>AES</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>CORT</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>PAX6</td>
<td>4</td>
<td>51</td>
</tr>
<tr>
<td>NFIC</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>AVPR2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Network-based prognosis - Summary

- A network-based method for biomarker discovery and cancer classification
  - Identifies marker based on connectivity
  - Classifies cancers based on global similarity of subnetwork activity
  - Significantly improves classification accuracy
  - **Able to discover key genes/pathways not identifiable with gene-focusing method**

Ongoing efforts in the lab

- Personalized cancer prognosis models
  - Further optimize the set of base classifiers
  - Generalize the idea to utilize multiple types of omics data
    - How to select the best omics data for different patients?
  - Combine patient-patient networks and gene-gene networks

- Network-based knowledge discovery to help understand high-throughput experimental results
  - Chromatin interaction
  - Pathway enrichment analysis

- TFBS binding using DNA 3D structure and multiple instance learning
Acknowledgements