Cancer subtype identification using large-scale omics data analysis

PhD public lecture

25.9.2019
Dvir Netanely
Ron Shamir’s Group
Tel Aviv University
STAR TREK: TNG 1987-1994
The vision of personalized medicine
STAR TREK: TNG 1987-1994
The vision of personalized medicine
The vision of personalized cancer medicine

**Current Medicine**
One Treatment Fits All

- Cancer patients with e.g. colon cancer
- Therapy
- Effect
- No effect
- Adverse effects

**Future Medicine**
More Personalized Diagnostics

- Cancer patients with e.g. colon cancer
- Blood, DNA, Urine and Tissue Analysis
- Therapy
- Effect
Algorithms

Statistics

Software

Big Data

Biology

Medicine
Talk Outline

• Introduction
  • What is cancer?
  • Cancer subtypes
  • Multi-omics data

• Improving Breast Cancer classification

• Improving Skin Cancer classification

• PROMO - A software tool for multi-omic data analysis and cancer subtyping
What is Cancer?

- Cancer is the second leading cause of death worldwide.
- It is characterized by uncontrolled cell proliferation, with the potential to invade or spread to other parts of the body (metastasis).
- Surgical removal of tumors, systemic chemotherapy and radiation therapy are still the main means to eradicate tumors today.
- However, improving our understanding of cancer biology and heterogeneity may translate into improved diagnosis and treatment.
Stages of tumor development

- Genetically altered epithelial cell
  - Cell divides more rapidly than normal

- Hyperplasia
  - Cells change form

- Dysplasia
  - Cells stay in one place

- In situ cancer
  - Cancer cells invade normal tissue and enter blood and lymph
  - Metastases form at distant sites

- Malignant tumor (cancer)
  - Metastases

The hallmarks of cancer


Cancer is very heterogeneous

- Cancer is challenging to treat efficiently because it is very heterogeneous
- Even tumors originating from the same organ can greatly vary
  - Genetic characteristics
  - Underlying biological mechanism
  - Survival risk
  - Response to treatment
Cancer subtypes

• One way to address the challenge of tumor heterogenicity is based on the ‘divide and conquer’ approach.

• Tumors are divided into homogenous groups based on similarity:
  • Pathological
  • Receptor status
  • Driver Mutation
  • Molecular profile (Genomics)

• Each group can then be characterized in terms of underlying biological mechanism, prognosis, response to treatment, targets for subtype-specific drug development.
But what exactly is a subtype?

- When does a group of tumors represent a subtype?
- How many subtypes are there?
Comparing subtype classifications
So, how many subtypes did you see?

- Potential cancer-subtype characteristics:
  - Well-defined biologically
  - Clinically-significant
  - Large enough
Cancer research in the multi-omic era

- Large cancer databases like TCGA (The Cancer Genome Atlas), provide multi-omic profiles for thousands of tumor samples, in addition to their associated clinical information.
- Tumorigenesis involves loss of cell regulation in various biological levels and therefore it is evident in various high throughput data types.
Research Goal

Improve cancer subtype classifications using large-scale multi-omic cancer datasets to promote personalized cancer medicine
Improving Breast Cancer classification

Dvir Netanely, Ayelet Avraham, Adit Ben-Baruch, Ella Evron, Ron Shamir

Breast Cancer

• The most **common** cancer among women
  • 1,300,000 cases, 450,000 deaths each year worldwide

• **Heterogeneous** with respect to
  • *Molecular alterations*
    • Several different cellular scenarios can lead to cancer
  • *Cellular composition*
    • Tumors are composed of several types of cells
  • *Clinical outcome*
    • Outcome prediction is challenging
Breast Cancer subtypes

ER+        Luminal     Luminal A
ER-        Luminal B   Basal
        Her2            Her2        Basal
        Luminal A
        Luminal B
Breast cancer subtypes

- Originally, therapeutic decisions were guided by parameters like tumor size, presence of lymph-node metastases and histological grade.
- In addition, the status of three biomarkers allowed the development of targeted therapies and proved predictive of treatment response:
  - Estrogen receptor (ER)
  - Progesterone receptor (PR)
  - Epidermal growth factor receptor 2 (HER2/ERBB2)

<table>
<thead>
<tr>
<th>Breast Cancer Subtypes</th>
<th>ER</th>
<th>PR</th>
<th>Her-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal A</td>
<td>+/−</td>
<td>+/−</td>
<td>−</td>
</tr>
<tr>
<td>Luminal B</td>
<td>+/−</td>
<td>+/−</td>
<td>+</td>
</tr>
<tr>
<td>Her-2 +</td>
<td>−</td>
<td>−</td>
<td>+</td>
</tr>
<tr>
<td>Triple Negative</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>
Breast cancer molecular subtypes

With the introduction of microarrays, gene expression profiling was used to define and characterize breast cancer subtypes.

"Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", Sørlie et al. 2001
The PAM50 Classifier (Parker, 2009)

- The PAM50 classifier can classify the molecular subtypes of breast cancer based on the expression levels of 50 genes.
- Highly used to this day.

Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes, Parker at al. 2009
TCGA’s Breast Cancer dataset

- TCGA’s database contains hundreds of breast tumor samples measured using 6 different high-throughput technologies.
- More than 150 clinical labels are available for each patient (such as age, race, tumor size, PAM50 label)
- Can we improve the classification of breast cancer tumor into clinically relevant subgroups?

“Comprehensive molecular portraits of human breast tumours”
The Cancer Genome Atlas Network
Rough beginning...

• Challenges
  • Obtaining TCGA’s data
  • Understanding TCGA’s coded clinical labels
  • Data kept updating as new samples were added regularly
  • Lack of analysis and visualization tools supporting large number of samples
  • Method uncertainty - Is it okay to use t-test on count data?
  • Couldn’t reproduce PAM50 partition
  • Interpreting the results biologically, medically
Using label color-bar in Expander
Feature-dynamic clustering with varying gene sets

PCA plots for running example 1: Four iterations over K-Means with K=5, 1000 genes per iteration
1148 Breast samples (Normal + Tumor)

Global unsupervised analysis on RNA-Seq (gene expression) data showed moderate agreement with PAM50 labels.
737 Luminal breast samples
Unsupervised analysis on RNA-Seq (gene expression)

Applying the K-Means algorithm on the 737 Luminal samples using K=2 splits the samples into two subgroups exhibiting better five-year prognostic value than the PAM50's Luminal-A/Luminal-B partition.
534 Luminal-A breast samples
Unsupervised analysis on RNA-Seq (gene expression)

- Reduced risk for 5-year recurrence
- Enriched for Lobular histology
- Lower proliferation rate
- Over expression of immune related genes
Updated survival statistics (2018)
LumA-R2 over-expressed genes in the T Cell receptor signaling pathway
The main distinguishing characteristics between the Luminal-A subgroups LumA-R1 and LumA-R2

<table>
<thead>
<tr>
<th>Group Characteristic</th>
<th>LumA-R1</th>
<th>LumA-R2</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recurrence free survival</td>
<td>Increased recurrence</td>
<td>Reduced recurrence</td>
<td>7.6e-3</td>
</tr>
<tr>
<td><strong>Histology</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrichment p-values for each group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ductal</td>
<td>(p=2.1e-05)</td>
<td>Lobular</td>
<td>(p=9.7e-12)</td>
</tr>
<tr>
<td>Age average</td>
<td>61.5</td>
<td>57.4</td>
<td>2.6e-05</td>
</tr>
<tr>
<td>Proliferation score</td>
<td>-0.4</td>
<td>-0.6</td>
<td>8.9e-25</td>
</tr>
<tr>
<td>Gene over-expression</td>
<td>194</td>
<td>1068</td>
<td></td>
</tr>
</tbody>
</table>

Out of 2000 genes used for clustering
The most enriched functional categories among the 1000 genes most differentially expressed between LumA-R1 and LumA-R2 samples.

<table>
<thead>
<tr>
<th>Enrichment Type</th>
<th>Term</th>
<th>#Genes</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene Ontology</strong></td>
<td>regulation of immune system process</td>
<td>152</td>
<td>3.74E-50</td>
</tr>
<tr>
<td></td>
<td>immune system process</td>
<td>201</td>
<td>3.65E-47</td>
</tr>
<tr>
<td></td>
<td>regulation of leukocyte activation</td>
<td>71</td>
<td>2.37E-28</td>
</tr>
<tr>
<td></td>
<td>regulation of multicellular organismal process</td>
<td>183</td>
<td>2.89E-28</td>
</tr>
<tr>
<td></td>
<td>cell activation</td>
<td>91</td>
<td>4.59E-28</td>
</tr>
<tr>
<td></td>
<td>regulation of response to external</td>
<td>73</td>
<td>8.18E-27</td>
</tr>
<tr>
<td></td>
<td>regulation of biological quality</td>
<td>218</td>
<td>1.82E-26</td>
</tr>
<tr>
<td></td>
<td>leukocyte activation</td>
<td>67</td>
<td>1.95E-26</td>
</tr>
<tr>
<td></td>
<td>positive regulation of cell activation</td>
<td>56</td>
<td>5.13E-24</td>
</tr>
<tr>
<td></td>
<td>T cell activation</td>
<td>45</td>
<td>4.93E-22</td>
</tr>
<tr>
<td></td>
<td>regulation of cell proliferation</td>
<td>128</td>
<td>1.83E-21</td>
</tr>
<tr>
<td><strong>KEGG Pathways</strong></td>
<td>Cytokine-cytokine receptor interaction</td>
<td>56</td>
<td>4.76E-22</td>
</tr>
<tr>
<td></td>
<td>Hematopoietic cell lineage</td>
<td>29</td>
<td>1.50E-17</td>
</tr>
<tr>
<td></td>
<td>Cell adhesion molecules (CAMs)</td>
<td>30</td>
<td>4.08E-13</td>
</tr>
<tr>
<td></td>
<td>Primary immunodeficiency</td>
<td>16</td>
<td>8.70E-13</td>
</tr>
<tr>
<td></td>
<td>Chemokine signaling pathway</td>
<td>31</td>
<td>1.14E-09</td>
</tr>
<tr>
<td></td>
<td>Complement and coagulation cascades</td>
<td>17</td>
<td>1.36E-08</td>
</tr>
<tr>
<td></td>
<td>T cell receptor signaling pathway</td>
<td>20</td>
<td>1.30E-07</td>
</tr>
<tr>
<td><strong>Wiki-Pathways</strong></td>
<td>TCR Signaling Pathway</td>
<td>10</td>
<td>1.55E-09</td>
</tr>
<tr>
<td></td>
<td>B Cell Receptor Signaling Pathway</td>
<td>10</td>
<td>1.72E-06</td>
</tr>
<tr>
<td></td>
<td>Focal Adhesion</td>
<td>11</td>
<td>5.88E-05</td>
</tr>
<tr>
<td></td>
<td>Complement Activation, Classical Pathway</td>
<td>6</td>
<td>8.38E-05</td>
</tr>
</tbody>
</table>
Section summary:
Unsupervised analysis of RNA-Seq expression data identified a Luminal-A subgroup associated with reduced recurrence (improved survival) and over expression of immune system genes.
Workflow for subtype detection

- Sample filtering
- Feature selection
- Preprocessing (flooring, ceiling, normalization)
- Sample clustering
- Label enrichment analysis
- Comparison to other partitions

Survival Analysis
- Gene clustering
- DEGs
- Gene enrichment analysis

Kaplan-Meier estimate of survival functions - Survival (7 day threshold)
**DNA-Methylation Data**

- Illumina’s bead chip array can measure the methylation level of over 450,000 CpG loci along the genome.
- Methylation at the beginning of a gene represses its expression.
Unsupervised analysis of breast tumors based on DNA Methylation data

All Tumor Samples

Luminal-A Samples

LumA-M1: Poorer 5-year survival, Hyper-methylation of developmental genes
Conclusions

- Our gene expression analysis identified a Luminal-A subgroup associated with increased risk for recurrence.
- Our DNA methylation data identified a Luminal-A subgroup associated with poorer survival.
- Cox multivariate analysis confirmed the prognostic value of the two signatures.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survival</th>
<th>Recurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
<td>p-value</td>
</tr>
<tr>
<td>LumA-R (1 vs 2)</td>
<td>0.56</td>
<td>0.369</td>
</tr>
<tr>
<td>LumA-M (2,3 vs 1)</td>
<td>6.68</td>
<td>0.004</td>
</tr>
<tr>
<td>Age (&lt;60 vs.&gt;=60 years)</td>
<td>11.2</td>
<td>0.003</td>
</tr>
<tr>
<td>Pathologic stage (I,II vs. III,IV)</td>
<td>2.12</td>
<td>0.255</td>
</tr>
<tr>
<td>ER Status</td>
<td>0.47</td>
<td>0.500</td>
</tr>
<tr>
<td>PR Status</td>
<td>1.48</td>
<td>0.726</td>
</tr>
<tr>
<td>Her2 Status</td>
<td>0.56</td>
<td>0.369</td>
</tr>
</tbody>
</table>

Integrating multi-omic datasets

- **Challenging** because datasets may differ in:
  - Value distribution
  - Number of features
  - Represent different biological entities (Gene, CpG, miRNA)
Integrating multi-omic datasets

Joint Multi-Omic Clustering

Expression

DNA Methylation

Compound Data Matrix

$D_{st} = \frac{d_{st}^{Exp} + d_{st}^{Meth}}{2}$
Multi-Omic Consensus Clustering
Integrating multi-omic datasets

Multi-Omic Correlation Graph

Expression

DNA Methylation
Workflow for subtype detection

- Sample filtering
- Feature selection
- Preprocessing (flooring, ceiling, normalization)
- Sample clustering
- Gene clustering
  - DEGs
- Gene enrichment analysis
- Label enrichment analysis
- Comparison to other partitions
- Survival Analysis
Improving Skin Cancer classification

Dvir Netanely, Neta Stern, Roma Parikh, Stav Leibo, Sarah Amar, Ronen Brenner, Carmit Levy, Ron Shamir
Cutaneous Melanoma

• Melanoma arises from the occurrence of genetic mutations in melanocytes, the pigment producing cells, which can be found in the skin, eye, inner ear, and leptomeninges.

• Cutaneous malignant melanoma represents the most aggressive and the deadliest form of skin cancer.
TCGA’s melanoma dataset

- Contains 474 samples
  - 104 Primary
  - 365 Metastatic
TCGA’s melanoma analysis (2015)

• TCGA’s analysis on 333 melanoma expression profiles, identified 3 prognostic groups:
  • Immune (n = 168; 51%)
  • Keratin (n = 102; 31%)
  • MITF-low (n = 59; 18%)

Unsupervised analysis of 469 melanoma expression profiles identifies 4 distinct subgroups
Characterization of the 4 sample groups

- Pigmentation / Melanosome membrane
- Cornification
- Immune

**Survival Analysis (1780 day threshold)**

- Line 1: \( p=0.000059, n=164 \)
- Line 2: \( p=0.000067, n=68 \)
- Line 3: \( p=0.2097, n=113 \)
- Line 4: \( p=0.0025, n=116 \)
### Main characteristics of the 4 sample-clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster name</th>
<th>TCGA Transcriptomic subtype enrichment</th>
<th>Survival</th>
<th>Tumor tissue type enrichment</th>
<th>Gene ontology enrichment of preferentially expressed genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Immune</td>
<td>Immune</td>
<td>Best</td>
<td>Regional lymph node</td>
<td>Immune response</td>
</tr>
<tr>
<td>2</td>
<td>Keratin/Primary-enriched</td>
<td>Keratin</td>
<td>Worst</td>
<td>Primary</td>
<td>Cornification</td>
</tr>
<tr>
<td>3</td>
<td>MITF-Low/Pigmentation-Low</td>
<td>MITF-Low</td>
<td>Good</td>
<td></td>
<td>Nervous system development</td>
</tr>
<tr>
<td>4</td>
<td>Pigmentation-High</td>
<td>Keratin</td>
<td>Bad</td>
<td></td>
<td>Melanin biosynthetic process</td>
</tr>
</tbody>
</table>
Sample cluster 4 is over-expressing melanosome related genes

<table>
<thead>
<tr>
<th>Set</th>
<th>Enriched with</th>
<th>#genes</th>
<th>Raw p-value</th>
<th>Empirical p-Value</th>
<th>Frequency in set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 neurogenesis - GO:0022008</td>
<td></td>
<td>68</td>
<td>7.10E-14</td>
<td>2.00E-04</td>
<td>19</td>
</tr>
<tr>
<td>1 somatodendritic compartment - GO:0036477</td>
<td></td>
<td>34</td>
<td>1.60E-08</td>
<td>4.00E-04</td>
<td>9.5</td>
</tr>
<tr>
<td>1 melanosome membrane - GO:0033162</td>
<td></td>
<td>6</td>
<td>2.34E-08</td>
<td>4.00E-04</td>
<td>1.7</td>
</tr>
<tr>
<td>1 cell body - GO:0044297</td>
<td></td>
<td>28</td>
<td>2.70E-08</td>
<td>6.00E-04</td>
<td>7.8</td>
</tr>
<tr>
<td>1 secondary metabolic process - GO:0019748</td>
<td></td>
<td>9</td>
<td>4.43E-08</td>
<td>6.00E-04</td>
<td>2.5</td>
</tr>
<tr>
<td>1 anatomical structure morphogenesis - GO:0009653</td>
<td></td>
<td>66</td>
<td>4.48E-08</td>
<td>6.00E-04</td>
<td>18</td>
</tr>
<tr>
<td>1 neuron projection development - GO:0031175</td>
<td></td>
<td>31</td>
<td>4.80E-08</td>
<td>6.00E-04</td>
<td>8.6</td>
</tr>
<tr>
<td>1 central nervous system development - GO:0007417</td>
<td></td>
<td>39</td>
<td>4.91E-08</td>
<td>6.00E-04</td>
<td>11</td>
</tr>
<tr>
<td>1 melanin metabolic process - GO:0006582</td>
<td></td>
<td>6</td>
<td>7.19E-08</td>
<td>8.00E-04</td>
<td>1.7</td>
</tr>
<tr>
<td>1 developmental pigmentation - GO:0048066</td>
<td></td>
<td>8</td>
<td>2.97E-07</td>
<td>0.0034</td>
<td>2.2</td>
</tr>
<tr>
<td>1 cellular component morphogenesis - GO:0032989</td>
<td></td>
<td>32</td>
<td>4.68E-07</td>
<td>0.005</td>
<td>8.9</td>
</tr>
<tr>
<td>1 brain development - GO:0007420</td>
<td></td>
<td>31</td>
<td>5.67E-07</td>
<td>0.0056</td>
<td>8.6</td>
</tr>
</tbody>
</table>

Melanosomes are tissue-specific organelles of pigment cells in which melanins are synthesized and stored.
Melanin related genes are prognostic biomarkers
Experimental validation:
Melanosomes diffuse outward from primary melanoma along a gradient
Experimental validation:
Melanosomes diffuse outward from metastatic melanomas
A 3-gene melanoma subtype classifier
A 3-gene melanoma subtype classifier
Conclusions

• Our analysis of TCGA’s melanoma dataset identified 4 prognostic subtypes.
• The high-pigmentation subgroup conferred poor 5-year survival, and may be associated with higher diffusion of melanosomes.
• We suggest a simple 3-gene classifier for predicting melanoma subtype.
PROMO
PROfiler of Multi-Omics data
A software tool for multi-omic data analysis and cancer subtyping

Dvir Netanely, Neta Stern, Itay Laufer, Ron Shamir
PROMO: An interactive tool for analyzing clinically-labeled multi-omic cancer datasets. bioRxiv 2019, Submitted to BMC Bioinformatics
Samples clustering and comparison to clinical labels

Histogram >SERIES PAM50 mRNA

Histogram >SERIES PAM50 mRNA

Histogram >SERIES PAM50 mRNA

PAM50

Normal

Her2

LuminalA

LuminalB

ER

Negative

Positive

PR

Negative

Positive

HER2

Negative

Positive

Normal

Tumor
What is PROMO?

• **PROMO** *(Profiler of Multi-Omics data)* is an interactive software for analyzing large genomic datasets of disease patients together with their associated clinical information.

• **PROMO** provides a single integrated solution for swiftly analyzing genomic data without writing a single line of code, and can, therefore, make the analysis of these data much easier for biologists and biomedical researchers.

• Supports **large** datasets containing thousands of samples.
• Provides tools for analyzing **clinical information**.
• Designed and built in the context of **cancer** classification.
• Support integrative **multi-omic** analyses.
Main Features

• Main features:
  • Import of data in various file formats
  • Preprocessing
  • Data exploration and visualization
  • Clustering of both samples and features
  • Supervised analysis using various statistical tools
  • Clinical label analysis
  • Survival analysis
  • Classification
  • Integrative multi-omic analysis
Figure 2 - Outline of PROMO’s subtype detection workflow

(1) Data import
- Text files
- UCSC XENA files
- Multi-omic Dataset collection files

(2) Preprocessing
- Label based sample filtering
- Variability feature filtering
- Gene symbol based feature filtering
- Normalization

(3) Exploration, Visualization
- Distribution plots
- Sorting
- PCA
- Survival analysis

(4) Subtype discovery
- Sample and feature clustering
- Multi-label enrichment analysis
- GO enrichment analysis

(5) Biomarker discovery
- Supervised tests
- Label based sample partitioning

(6) Classification
- Decision tree generation
- Cross validation

(7) Data export
- Text files
- Multi-label expression matrix
Main Features

Data Import
- Tabular text file
- GEO Series file (Direct download)
- TCGA data from UCSC XENA

Preprocessing
- Flooring, Ceiling, Log2 transformations
- Variability and range filters
- Filter samples by label

Data Exploration and Visualization
- Distribution and variability plots
- PCA and t-SNE
- Sorting

Clustering
- K-Means
- Hierarchical clustering
- Click

Enrichment Analysis & Partition Comparison
- Clinical label enrichment analysis on sample clusters
- Gene ontology enrichment analysis on gene clusters
- Comparison of sample partitions using HG and Chi-Square tests

Supervised tests
- Identification of differentially expressed genes (T-test, Rank-sum test, ANOVA, FDR)

Survival Analysis
- Kaplan-Meier plots
- Log-rank survival test

Classification
- Automatic decision tree generation for selected label

Multi-Omics
- Multi-omic dataset integration
- Joint multi-omic sample clustering: Consensus clustering, SNF, NEMO
- Identification of inter-omic feature correlations
Summary

• Large-scale multi-omic cancer datasets are expected to play an important role in deciphering cancer biology and pathology.

• In this work we have integrated omics data with clinical information for improving the classification of breast and skin cancer.

• Our methodology was implemented in PROMO, and can be applied to many other cancer types as well as to future large genomic datasets.
THANKS

- Prof. Ron Shamir
- Past and present members of the Shamir group
- Gilit Zohar Oren

**Breast Cancer Project**
- Ella Evron, Ayelet Avraham, Adit Ben Baruch

**Skin Cancer Project**
- Carmit Levy and her group
- Ronen Brenner

**PROMO**
- Neta Stern, Itay Laufer
THANK YOU