TCGA breast cancer study

Towards a breast cancer GRN ...

Motifs in GRNs ... TMmiR
The Cancer Genome Atlas (TCGA): breast cancer

TCGA consortium analysed primary breast cancers by
- genomic DNA copy number arrays,
- DNA methylation,
- exome sequencing,
- messenger RNA arrays,
- microRNA sequencing and
- reverse-phase protein arrays.

Combining data from 5 platforms showed the existence of 4 main breast cancer classes. Each of them shows significant molecular heterogeneity.

Somatic mutations in only 3 genes (*TP53*, *PIK3CA* and *GATA3*) occurred at >10% incidence across all breast cancers.
Breast cancer genomes in TCGA

Tumour samples are grouped by mRNA subtype: luminal A ($n = 225$), luminal B ($n = 126$), HER2E ($n = 57$) and basal-like ($n = 93$). Clinical features: dark grey, positive or T2–4; white, negative or T1; light grey, N/A or equivocal. N, node status; T, tumour size.

Right: significantly mutated genes with frequent copy number amplifications (red) or deletions (blue). Far-right: non-silent mutation rate per tumour (mutations per megabase, adjusted for coverage).

RegulonDB: database with information on transcriptional regulation and operon organization in \textit{E.coli}; 105 regulators affecting 749 genes

→ 7 regulatory proteins (CRP, FNR, IHF, FIS, ArcA, NarL and Lrp) are sufficient to directly modulate the expression of more than half of all \textit{E.coli} genes.

Martinez-Antonio, Collado-Vides, Curr Opin Microbiol 6, 482 (2003)
Review (bioinformatics III) – Regulatory cascades in *E. coli*

When more than 1 TF regulates a gene, the order of their binding sites is as given in the figure.

Arrowheads indicate positive regulation when the position of the binding site is known.

Horizontal bars indicate negative regulation when the position of the binding site is known.

In cases where only the nature of regulation is known, without binding site information, + and – are used to indicate positive and negative regulation.

The DNA binding domain families are indicated by circles of different colours.

The names of global regulators are in bold.

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)
Aim: construct GRN for breast cancerogenesis

We found
- 1317 differentially expressed genes,
- 2623 differentially methylated genes,
- 121 differentially expressed miRNAs between 131 tumor and 20 normal tissues.

Organize genes into modules

The expression profiles of the 1317 identified differentially expressed genes were used to compute the co-regulation strength between genes.

An undirected co-expression network was obtained by hierarchical clustering (HCL).

HCL yielded 10 segregated network modules that contain between 26 and 295 gene members (different colors).
**Organize genes into modules**

<table>
<thead>
<tr>
<th>Module</th>
<th>Gene count</th>
<th>Top GO category</th>
<th>Top KEGG categories</th>
<th>Key driver count</th>
<th>Key drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>black</td>
<td>41</td>
<td>Regulation of transcription</td>
<td>Pathways in cancer, Renal cell carcinoma</td>
<td>5</td>
<td>SORBS3, ZNF43, ZNF681, RBMX, POU2F1</td>
</tr>
<tr>
<td>blue</td>
<td>247</td>
<td>Nucleobase, nucleoside, nucleotide and</td>
<td>Cell cycle, Prostate cancer, Melanoma</td>
<td>9</td>
<td>AR, BRCA1, ESR1, JUN, MYB, RPN1, E2F1, E2F2, PPARD</td>
</tr>
<tr>
<td>brown</td>
<td>195</td>
<td>Anatomical structure morphogenesis</td>
<td>Leukocyte transendothelial migration</td>
<td>5</td>
<td>TMOD3, CREB1, POU5F1, SP3, TERT</td>
</tr>
<tr>
<td>green</td>
<td>110</td>
<td>Cellular macromolecule metabolic</td>
<td>Endometrial cancer, Insulin signaling pathway</td>
<td>15</td>
<td>B4GALT7, OS9, CDC34, MAN2C1, MYO1C, SH3GLB2, INPP5E, PLXNB1, USF2, PPP1R12C, CDK9, DAP, E4F1, E2F4, USF1</td>
</tr>
<tr>
<td>grey</td>
<td>148</td>
<td>Anatomical structure development</td>
<td>Sulfur metabolism</td>
<td>18</td>
<td>AHCTF1, NQO2, FGFR2, CCDC130, ABCG4, BIRC6, CA6, SP4, RNF2, SPPR1B, C16orf65, DNAJC5G, SNCAIP, GRIK5, SLC6A4, SMAD1, DAD1, DAD1, POU4F2</td>
</tr>
<tr>
<td>magenta</td>
<td>26</td>
<td>Regulation of metabolic process</td>
<td>p53 signaling pathway, Alzheimer’s disease</td>
<td>3</td>
<td>ATF6, NGEF, POGK</td>
</tr>
<tr>
<td>pink</td>
<td>30</td>
<td>Transcription initiation from RANA</td>
<td>Basal transcription factors</td>
<td>4</td>
<td>CCDC92, TMEM70, RNF139, E2F5</td>
</tr>
<tr>
<td>red</td>
<td>93</td>
<td>Regulation of cellular process</td>
<td>Endometrial cancer, Neurotrophin signaling pathway</td>
<td>14</td>
<td>ATP1B1, STAT3, ABCB8, MYC, TGFBR1, SP1, TP53, PGCF1, SUMF2, GTF3A, IPO13, GMPPA, HTR5, TGIF1</td>
</tr>
<tr>
<td>turquoise</td>
<td>295</td>
<td>Regulation of cellular metabolic</td>
<td>p53 signaling pathway, Pancreatic cancer, Apoptosis</td>
<td>2</td>
<td>UBL5, RNF111</td>
</tr>
<tr>
<td>yellow</td>
<td>132</td>
<td>Immune system process</td>
<td>Chemokine signaling pathway, Natural killer cell mediated cytotoxicity</td>
<td>19</td>
<td>APOC1, CD2, CD79B, LRRC28, DAPK1, FAM124B, EML2, LAP3, TSPAN2, FCRL3, ELM01, SLC7A7, RASSF5, SLC31A2, TRAF3IP3, GALNT12, ITGA4, SP1, TFA2A</td>
</tr>
</tbody>
</table>

Enriched GO categories and KEGG pathways determined among genes in module by hypergeometric test (p < 0.05).

Key driver genes

Key regulators (drivers) in the constructed modules were identified by determining the minimal set of nodes that regulate the entire module.

These nodes typically include the nodes with highest degree plus some further ones that are required to „reach“ the remaining target genes.

Single stranded miRNAs are incorporated into the RISC complex.

This complex then targets the miRNA e.g. to the target 3’ untranslated region of a mRNA sequence to facilitate repression and cleavage.

AA, poly A tail; m7G, 7-methylguanosine cap; ORF, open reading frame.
Binding partners of miRNAs

Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules.

solution NMR-structure of let-7 miRNA: lin-41 mRNA complex from C. elegans

The main function of miRNAs is to down-regulate translation of their target mRNAs.

miRNAs typically have incomplete base pairing to a target and inhibit the translation of many different mRNAs with similar sequences.

In contrast, siRNAs typically base-pair perfectly and induce mRNA cleavage only in a single, specific target.
discovery of let7

The first two known microRNAs, lin-4 and let-7, were originally discovered in the nematode *C. elegans*.

They control the timing of stem-cell division and differentiation. let-7 was subsequently found as the first known human miRNA.

let-7 and its family members are **highly conserved** across species in sequence and function. Misregulation of let-7 leads to a less differentiated cellular state and the development of cell-based diseases such as cancer.

www.wikipedia.org
Construct regulatory interactions

* For the 7 smallest modules, we collected the related directed regulatory interactions available in 3 online regulatory databases (JASPAR, TRED, MsigDB).

These were used as a prior for a Bayesian learner to learn the causal probabilistic regulatory interactions and to generate a directed network topology.

* We removed 89 inferred interactions whose target genes are downregulated and their expression profiles showed absolute anti-correlation measure > 0.65 with their methylation profiles.

In those cases we reasoned that downregulation of these target genes was most likely due to their promoter methylation and not due to TF binding.
Construct regulatory interactions involving miRNAs

* For the set of differentially expressed miRNAs, which were either up- or down-regulated between the tumor and normal samples, we used miRTrail via MicroCosm Targets V5 to extract their target mRNAs (regulated genes) and overlapped them with the identified differentially expressed mRNAs.

* We used the experimentally validated database TransmiR to retrieve the regulatory genes (TFs) that potentially regulate the differentially expressed miRNAs.

### miRNA-mRNA interactions in breast cancer network

<table>
<thead>
<tr>
<th>miRNA-mRNA interactions</th>
<th>Genes</th>
<th>Top GO category</th>
<th>Top KEGG categories</th>
<th>Key driver count</th>
<th>Key drivers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>869</td>
<td>Regulation of macromolecule metabolic process</td>
<td>Pathways in cancer, Pancreatic cancer, Prostate cancer</td>
<td>17</td>
<td>MYC, ATG4C, TGFβ1, NFkB1, AKT1, EGR1, TP53, SOX10, SPI1, MECP2, E2F3, CREB1, TCF3, TPP1, FLICE, LPS, PACS1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>miRNAs</th>
<th>miRNA count</th>
<th>Top functional categories</th>
<th>Top HMDD categories</th>
<th>Key driver count</th>
<th>Key drivers</th>
</tr>
</thead>
</table>

For the 10 gene modules identified in TF-mRNA interactions, we list counts of the involved genes, the most significant GO and KEGG terms, and the identified key driver genes from each module. Similarly for the miRNA-mRNA interactions, we list the key driver molecules of both genes and miRNAs. The driver genes, whose protein products are known to be targeted by drugs, are in bold.

GRN modules

Gene network modules of TF-gene interactions. (a) Topological overlap matrix (TOM) heatmap corresponding to the 10 co-expression modules. Each row and column of the heatmap represent a single gene. Spots with bright colors denote weak interaction whereas darker colors denote strong interaction. Dendrograms on the upper and left sides show the hierarchical clustering tree of genes. (b), (c), and (d) are the final GRN networks highlighting the identified key drivers genes for the green, magenta, and red modules, respectively. Square nodes : identified driver genes that are targeted by drugs.

SS 2015 - lecture 12

Modeling Cell Fate

Regulatory interactions of the 17 key driver genes identified from miRNA-mRNA interactions.

Large nodes: key driver genes

Small nodes: miRNAs, which regulate or are regulated by these driver genes.

Square nodes: driver genes that are targeted by available drug molecules.

We identified 94 driver genes from the TF-mRNA interactions and 17 driver genes from the miRNA-mRNA interactions. 5 breast cancer associated genes $CREB1$, $MYC$, $TGFBI$, $TP53$, and $SPI1$ were common in both sets -> in total 106 driver genes.

31% (33 proteins) of the proteins belonging to the identified driver genes are binding targets of at least one anti-breast cancer drug -> validates our approach
TFmiR: identify regulatory motifs in GRNs

TFmiR web user interface

For each interaction type
- Venn diagram
- Basic statistics
- ORA analysis of genes and miRNAs (if any)
- Significance of overlaps with the related database

TF-gene
miRNA→gene
TF-miRNA
miRNA-miRNA

Combine all TF and miRNA co-regulatory interactions

Disease-specific network
disease-related genes and miRNAs

ORKA analysis
Network analysis
- Topological features
- Network visualization
- Key nodes / hot spot identification
- Significance and coverage rate of disease

TF-miRNA co-regulatory motifs

Functional similarity of co-targeted genes

## Data sources used for TFmiR

Table S1. The integrated databases and interaction types in TFmiR.

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Databases (P/E) *</th>
<th>Genes</th>
<th>miRNAs</th>
<th>Regulatory links</th>
<th>Version / frozen date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TF → gene</strong></td>
<td>TRANSFAC (E) (1)</td>
<td>1279</td>
<td>--</td>
<td>2943</td>
<td>V11.4</td>
</tr>
<tr>
<td></td>
<td>OregAnno (E)(2)</td>
<td>1132</td>
<td>--</td>
<td>1083</td>
<td>Nov 2010</td>
</tr>
<tr>
<td></td>
<td>TRED (P) (3)</td>
<td>3038</td>
<td>--</td>
<td>6462</td>
<td>2007</td>
</tr>
<tr>
<td><strong>TF → miRNA</strong></td>
<td>TransmiR (E) (4)</td>
<td>158</td>
<td>175</td>
<td>567</td>
<td>V1.2, Jan 2013</td>
</tr>
<tr>
<td></td>
<td>PMID20584335 (E) (5)</td>
<td>58</td>
<td>56</td>
<td>102</td>
<td>Apr 2009</td>
</tr>
<tr>
<td></td>
<td>ChipBase (P) (6)</td>
<td>119</td>
<td>1380</td>
<td>33087</td>
<td>V1.1, Nov 2012</td>
</tr>
<tr>
<td><strong>miRNA → gene</strong></td>
<td>miRTarBase (E)(7)</td>
<td>2244</td>
<td>551</td>
<td>5640</td>
<td>V4.5, Nov 2013</td>
</tr>
<tr>
<td></td>
<td>TarBase (E) (8)</td>
<td>422</td>
<td>79</td>
<td>492</td>
<td>V7.0</td>
</tr>
<tr>
<td></td>
<td>miRecords (E)(9)</td>
<td>543</td>
<td>157</td>
<td>780</td>
<td>Mar 2009</td>
</tr>
<tr>
<td></td>
<td>starBase (P)(10)</td>
<td>5720</td>
<td>249</td>
<td>56051</td>
<td>V2.0, Sept 2013</td>
</tr>
<tr>
<td><strong>miRNA → miRNA</strong></td>
<td>PmmR (P) (11)</td>
<td>312</td>
<td>3846</td>
<td></td>
<td>Mar 2011</td>
</tr>
</tbody>
</table>

* (P) means predicted interactions and (E) means experimentally validated interactions.
We postulated that a TF and a miRNA may act occasionally on the same gene (or its transcribed mRNA).

In motif (a), a TF could first stimulate gene expression.

Later, the miRNA would degrade the transcript.

-> search generated networks for such co-regulatory motifs

Statistical significance of motifs

To evaluate the significance of each FFL motif type, we compare how often they appear in the real network to the number of times they appear in randomized ensembles preserving the same node degrees.

We applied a degree preserving randomization algorithm.

Each random network was generated by \( 2 \times \) number of edges steps, where in each step we choose 2 edges \( e_1 = (v_1, v_2) \) and \( e_2 = (v_3, v_4) \) randomly from the network and swap their start and end nodes, i.e. \( e_3 = (v_1, v_4) \) \( e_4 = (v_3, v_2) \).

We construct 100 such random networks and count the motifs in them.

\( N_{\text{motif_random}} \) is the number of random networks that contain more than or equal numbers of a certain motif than the real network.

The we compute the p-value for the motif significance as \( p = N_{\text{motif_random}} / 100 \).
**Figure S4.** A composite FFL motif involves the TF *SPI1*, the miRNA *has-mir-155*, and the target gene *FLI1*. The co-regulated nodes are also visualized and are further tested whether they compose a cooperative functional module in breast cancerogenesis (see Fig S5).

Enriched biological functions in breast cancer network

Table S2. The most significant functions and diseases enriched in the miRNA nodes of the breast cancer disease network (12).

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>miRNAs Count</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function</td>
<td>Epithelial-mesenchymal transition</td>
<td>17</td>
<td>0.022</td>
</tr>
<tr>
<td>Function</td>
<td>glucose metabolism</td>
<td>4</td>
<td>0.048</td>
</tr>
<tr>
<td>Disease</td>
<td>Breast Neoplasms</td>
<td>67</td>
<td>1.43E-25</td>
</tr>
<tr>
<td>Disease</td>
<td>Lung Neoplasms</td>
<td>50</td>
<td>4.33E-17</td>
</tr>
<tr>
<td>Disease</td>
<td>Neoplasms</td>
<td>44</td>
<td>3.15E-15</td>
</tr>
<tr>
<td>Disease</td>
<td>Ovarian Neoplasms</td>
<td>43</td>
<td>1.30E-14</td>
</tr>
<tr>
<td>Disease</td>
<td>Adenocarcinoma</td>
<td>27</td>
<td>2.59E-13</td>
</tr>
<tr>
<td>Disease</td>
<td>Pancreatic Neoplasms</td>
<td>39</td>
<td>7.30E-13</td>
</tr>
<tr>
<td>Disease</td>
<td>Prostatic Neoplasms</td>
<td>41</td>
<td>3.49E-12</td>
</tr>
<tr>
<td>Disease</td>
<td>Melanoma</td>
<td>45</td>
<td>1.25E-11</td>
</tr>
<tr>
<td>Disease</td>
<td>Colonic Neoplasms</td>
<td>32</td>
<td>4.67E-11</td>
</tr>
<tr>
<td>Disease</td>
<td>Colorectal Neoplasms</td>
<td>45</td>
<td>5.69E-11</td>
</tr>
</tbody>
</table>

### Table S4. The identified key gene nodes in the breast cancer network (12) whose protein products are targeted by anti-cancer drugs. (1) means that at least one drug that targets this gene product is reported in this database, and (0) means no drugs are reported for the respective gene in this database. Not included are substances that are known to be cancerogenous or mutagenic.

<table>
<thead>
<tr>
<th>Target gene</th>
<th>Drug and antineoplastic agents</th>
<th>CTD</th>
<th>PharmGKB</th>
<th>Cancer Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT1</td>
<td>U 0126; tyrphostin AG 1478; Ursodeoxycholic Acid; Valproic Acid; tyrphostin AG 1024; trametinib; Tretinoin</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BRCA2</td>
<td>Tretinoin; trichostatin A; Estradiol; transplatin; troglitazone; Tunicamycin; fulvestrant</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ESR1</td>
<td>exemestane; tamoxifen</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TGFβ1</td>
<td>Doxorubicin; Fluorouracil; Thalidomide; Entinostat; Hyaluronidase</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TP53</td>
<td>4-biphenylmine; alliin; Apigenin; Atropine; Bicalutamide; Butylideneaphthalide</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

End of lecture

This concludes our small tour on modeling cell fate.

We looked at:
- Circadian clocks
- Cell cycle
- Cell differentiation
- Cancerogenesis

We showed you all the necessary techniques to generate GRNs by yourself.

Next week: mini-test 3, no lecture

Also open: assignment #6

If you like to conduct a Master thesis in these areas please contact me.