AMARETTO:
Multi-omics data fusion for cancer data

Magali Champion, K. Brennan, T. Croonenborghs, A. Gentles, N. Pochet, O. Gevaert

Séminaire MIAT
Motivation

Create mechanistic models of cancer to:

- Understand how gene expression is influenced by genomic events,
- Identify cancer driver genes and their targets.

Need to develop statistical methods that allow to integrate multi-omics data:

- Genomic (DNA copy number),
- Transcriptomic (gene expression, microARN),
- Methylomic (DNA methylation).

Extend these methods to a pancancer analysis.

· · · · · · · · · · Face the big data challenge · · · · · · · · · ·
Data overview

NIH project to extensively characterize the cancer genome (more than 20 cancers and 500 patients each)

- Gene & miRNA expression (Agilent & Affy microarray - RNA sequencing)
- Copy number (Affy SNP 6.0)
- DNA methylation (Agilent Infinium (27k))
- Mutation (DNA sequencing)
- Pathology images
- Medical images (MRI, CT)
## Data overview

<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>TCGA code</th>
<th>Samples</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bladder cancer</td>
<td>BLCA</td>
<td>181</td>
<td>15,432</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>BRCA</td>
<td>985</td>
<td>16,020</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>COADREAD</td>
<td>589</td>
<td>15,533</td>
</tr>
<tr>
<td>Glioblastoma</td>
<td>GBM</td>
<td>501</td>
<td>17,811</td>
</tr>
<tr>
<td>Head and Neck squamous carcinoma</td>
<td>HNSC</td>
<td>371</td>
<td>15,828</td>
</tr>
<tr>
<td>Kidney clear cell carcinoma</td>
<td>KIRC</td>
<td>509</td>
<td>16,123</td>
</tr>
<tr>
<td>Acute myeloid carcinoma</td>
<td>LAML</td>
<td>173</td>
<td>14,296</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>LUAD</td>
<td>489</td>
<td>16,092</td>
</tr>
<tr>
<td>Lung squamous carcinoma</td>
<td>LUSC</td>
<td>490</td>
<td>16,219</td>
</tr>
<tr>
<td>Ovarian cancer</td>
<td>OV</td>
<td>541</td>
<td>17,814</td>
</tr>
<tr>
<td>Endometrial cancer</td>
<td>UCEC</td>
<td>508</td>
<td>15,706</td>
</tr>
</tbody>
</table>
AMARETTO:
Multi-omics data fusion for cancer data

Discovering cancer driver genes and their targets
Method: AMARETTO algorithm

- Multi-omics data fusion of gene expression, copy number and DNA methylation

- Two-step algorithm:
  1. Identifying driver genes based on copy number and methylation,
  2. Associating cancer driver genes with their downstream targets.

Gevaert et al., Interface Focus, 2013
Step 1: Generating the list of candidate drivers

If gene expression can be explained by genomic events

\[ \text{Candidate driver gene} \]

Model gene expression as a function of copy number and DNA methylation:

\[ \text{Expression}_{\text{Gene}_i} = f(\beta_1 \text{Methylation}_{\text{Gene}_i} + \beta_2 \text{Copy Number}_{\text{Gene}_i}). \]

\[ \text{MethylMix} \quad \text{GISTIC} \]
Step 1 : Cancer driver gene filtering

Use dedicated modeling on copy number and DNA methylation before AMARETTO integration:
- GISTIC : identifies recurrent copy number alterations,
- MethylMix : identifies hyper & hypo-methylated genes.

Transfer of a methyl-group to the DNA:
- causes gene expression silencing,
- deregulated in cancer (hyper/hypo-methylation)
Step 1: Cancer driver gene filtering (MethylMix)

Remarks:
- No formal method to model hyper and hypo methylated in cancer
- The normal state is unknown

Step 1:
- Typical DNA methylation data distribution
- Beta value

Gevaert et al., Genome Biology, 2015
Step 1: Cancer driver gene filtering (MethylMix)

Step 2:
- mixture of beta distributions
- identification of two components
Step 1: Cancer driver gene filtering (MethylMix)

Step 3:
- comparison with DNA methylation in normal samples

An example of hyper-methylation of BRCA1 in ovarian cancer
Step 1: Cancer driver gene filtering (MethylMix)

Step 4:
- inverse correlation with gene expression

R-square statistic to quantify amount of variation explained
## MethylMix results

<table>
<thead>
<tr>
<th>TCGA code</th>
<th>Total of Genes</th>
<th>Hyper-methylated</th>
<th>Hypo-methylated</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLCA</td>
<td>15,432</td>
<td>443</td>
<td>74</td>
</tr>
<tr>
<td>BRCA</td>
<td>16,020</td>
<td>798</td>
<td>203</td>
</tr>
<tr>
<td>COADREAD</td>
<td>15,533</td>
<td>847</td>
<td>177</td>
</tr>
<tr>
<td>GBM</td>
<td>17,811</td>
<td>246</td>
<td>140</td>
</tr>
<tr>
<td>HNSC</td>
<td>15,828</td>
<td>728</td>
<td>101</td>
</tr>
<tr>
<td>KIRC</td>
<td>16,123</td>
<td>319</td>
<td>251</td>
</tr>
<tr>
<td>LAML</td>
<td>14,296</td>
<td>470</td>
<td>77</td>
</tr>
<tr>
<td>LUAD</td>
<td>16,092</td>
<td>576</td>
<td>182</td>
</tr>
<tr>
<td>LUSC</td>
<td>16,219</td>
<td>605</td>
<td>133</td>
</tr>
<tr>
<td>OV</td>
<td>17,814</td>
<td>234</td>
<td>229</td>
</tr>
<tr>
<td>UCEC</td>
<td>15,706</td>
<td>618</td>
<td>238</td>
</tr>
</tbody>
</table>
Step 1: Results for glioblastoma

The graph shows the number of genes for different variance of gene expression explained (R-square) categories. The categories are 20%, 30%, 40%, and 50%. The bars are color-coded as follows:

- **CNV**: Blue
- **MET**: Red
- **CNV and MET**: Green

- At 20%, there are 1137 CNV genes, 299 MET genes, and 229 CNV and MET genes.
- At 30%, there are 370 CNV genes, 116 MET genes, and 89 CNV and MET genes.
- At 40%, there are 133 CNV genes, 50 MET genes, and 32 CNV and MET genes.
- At 50%, there are 55 CNV genes, 15 MET genes, and 11 CNV and MET genes.
Step 2: Associating candidate drivers with their downstream targets

\[ \forall \text{ Module}_i, \quad \text{Expression}_{\text{Module}_i} = f(\alpha_1 \text{Driver}_1 + \ldots + \alpha_n \text{Driver}_n) \]

Linear regression + lasso regularization
Module network
### Results obtained for **Glioblastoma**

<table>
<thead>
<tr>
<th>Cancer driver</th>
<th>Number of modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZNF300</td>
<td>10</td>
</tr>
<tr>
<td>TNFRSF1A</td>
<td>10</td>
</tr>
<tr>
<td>PTRF</td>
<td>8</td>
</tr>
<tr>
<td>WWTR1</td>
<td>8</td>
</tr>
<tr>
<td>MYT1</td>
<td>7</td>
</tr>
<tr>
<td>PYCARD</td>
<td>7</td>
</tr>
<tr>
<td>PATZ1</td>
<td>7</td>
</tr>
<tr>
<td>BASP1</td>
<td>6</td>
</tr>
<tr>
<td>RAB32</td>
<td>6</td>
</tr>
<tr>
<td>SATB1</td>
<td>6</td>
</tr>
</tbody>
</table>

Top cancer drivers in GBM are:
- ZNF300, associated with immune system in Leukemia
- TNFRSF1A, associated with NF-kB pathway and angiogenesis in GBM
- RAB32, associated with hyper-methylation
AMARETTO modules capture pathways

Results obtained for **Ovarian cancer**
AMARETTO identifies drivers of subtypes

TCGA molecular subtypes of ovarian cancer:
- immuno-reactive
- differentiated
- mesenchymal
- proliferative

→ modules correlated with subtypes point to potential driver genes

Bell et al., Nature, 2011
AMARETTO identifies drivers of subtypes
To a pancancer AMARETTO analysis?

After running AMARETTO on the 11 cancer sites, 11 module networks were produced, with:

- 100 modules per network,
- an averaged number of 408 drivers per network,
- between 348 (BRCA) and 452 (LUAD) driver genes.

In addition,

- each module from all cancer sites is regulated by an averaged number of 7.67 drivers,
  → sparse method
  → most of them are methylated
- 45 drivers regulate more than 15 modules across all cancer sites.
To a pancancer AMARETTO analysis?
To a pancancer AMARETTO analysis?

<table>
<thead>
<tr>
<th>Gene</th>
<th>Number of regulated modules</th>
<th>Number of involved cancers</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSTL1</td>
<td>31</td>
<td>7</td>
</tr>
<tr>
<td>IFFO1</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>MLPH</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>SPARCL1</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>CLIP3</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>MFAP4</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>BEND5</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>NUAK1</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>CAPS</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>PP1R16B</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>OLFML1</td>
<td>23</td>
<td>5</td>
</tr>
<tr>
<td>SLA</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>DDR2</td>
<td>20</td>
<td>7</td>
</tr>
</tbody>
</table>
To a pancancer AMARETTO analysis?
AMARETTO:
Multi-omics data fusion for cancer data

Pancancer module networks
Pancancer analysis

AMARETTO

GBM modules

OV modules

GBM
HNSC
KIRC
LAML
LUAD
LUSC
OV
UCEC

BLCA
BRCA
COADREAD
Pancancer analysis

BLCA
BRCA
COADREAD
GBM
HNSC
KIRC
LAML
LUAD
LUSC
OV
UCEC

COADREAD modules
GBM modules
LAML modules
UCEC modules

BLCA modules
HNSC modules
LUAD modules
OV modules

BRCA modules
KIRC modules
LUSC modules

AMARETTO
Pancancer module network

**Hypergeometric test** to measure whether there is a significant association between all pairs of modules from all cancer types.

- **p-value** to compare modules
- **Enrichment score using negative log of the p-value**

![Diagram of Pancancer module network with module X and module Y comparisons and p-value representation.](image)
Pancancer module network

**Hypergeometric test** to measure whether there is a significant association between all pairs of modules from all cancer types.
Community detection algorithm

To detect communities, we used the Girvan Newman algorithm (edge betweenness detection algorithm), which consists in :

1- computing the betweenness score of all graph edges (numbers of shortest paths that run along each edge),
2- removing from the graph the edge with the highest score,
3- running Step 1 and Step 2 with the new graph obtained after Step 2.

⚠️ Weight edge betweenness score with the $- \log p$-value score.

Newman et al., Physical Review E., 2004
Pancancer AMARETTO results

Edge betweeness algorithm detected 20 communities

- between 9 and 74 modules
- averaged number of 30.5 modules
- around 10 cancer sites represented in each community
Pancancer histone community

- Contains 11 modules representing all different cancers (one module for each cancer)
- Overlapping cancer driver genes are part of histones
- Enrichment in cell cycle genes
Pancancer smoking community

- Contains 15 modules representing 8 different cancers (KIRC, GBM and LAML are not represented)

- Overlapping cancer driver genes
  - 3 genes in 3 modules
  - 1 gene in 8 modules
  - GPX2

- Enrichment in smoking related pathways
Collecting clinical data, GPX2 expression is significantly associated with smoking profile.
Pancancer smoking community

- Head and Neck cancer
  - Corr = 0.24
  - p-value = $e^{-6}$

- Lung adenocarcinoma
  - Corr = 0.42
  - p-value = $e^{-20}$

- Lung squamous cell carcinoma
  - Corr = 0.26
  - p-value = $e^{-9}$

- Endometrial cancer
  - Corr = 0.24
  - p-value = $e^{-8}$

Oxidative stress signature vs. GPX2 expression
Pancancer immune response community

- Contains 15 modules representing 10 different cancers (only KIRC is not represented here)

- Overlapping cancer driver genes
  - 6 genes in 4 modules
  - 1 gene in 6 modules
  - 1 gene in 10 modules

  **OAS2**

- Enrichment in immune response pathways
Pancancer immune response community

- Most of the drivers are part of interferons.

Type I and II:
AIM2, CCL5, EPSTI1, ETV7, GBP4, HCP5, HLA-F, IFI35, IRF7, ISG20, MX2, NMI, OAS2, PSMB8, RARRES3, SLC15A3, SP100, TMEM140, TRIM21, TRIM22, XAF1

Type I-II and III:
BAXT2, BST2, IFI16, OAS1, PARP9, SP110

Type I: 21
Type II: 21
Type III: 6
Conclusion

AMARETTO
- Identifies driver genes through multi-omics data integration
- Connects them to their downstream targets

Pancancer AMARETTO
- Identifies major oncogenic pathways and master regulators involved in multiple cancers
- Identifies an interferon master regulator involved in immune response pathway

AMARETTO extension
- Will allow the integration of miRNA data to identify drivers miRNAs and their effect on mRNAs

R-package available at https://bitbucket.org/gevaertlab/pancanceramaretto
Thanks for your attention!


C.H. Mermel et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology 12 :R41, 2011.
