Identification of Differentially Expressed Gene Modules in Heterogeneous Diseases

O. Zolotareva[1], S. Khakabimamaghani[2], O. Isaeva[3,4], M. Ester[2,5]

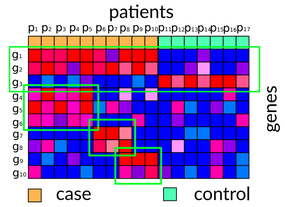
1. IRTG DiDy, Bielefeld University, Germany 2. School of Computing Science, Simon Fraser University, Canada
3. Center of Life Sciences, Skolkovo Institute of Science and Technology, Russia
4. BostonGene LLC, USA 5. Vancouver Prostate Centre, Canada.

Novel network-constrained **biclustering method** reveals differentially expressed gene modules in **breast cancer**

BACKGROUND

Disease heterogeneity

- $g_{1\text{-}3}$ are up-regulated in case group compared to controls
- $g_{4\text{-}6}$, $g_{7\text{-}8}$ and $g_{9\text{-}10}$ are up-regulated only in **subgroups** of cases
- disease **subgroups** may be **unknown**

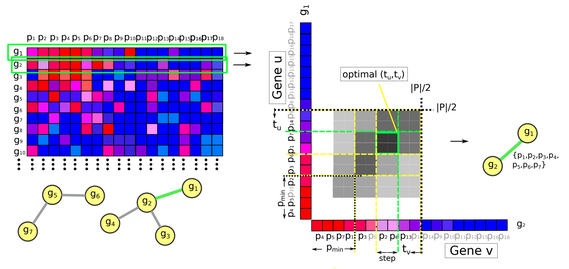


METHOD

DESMOND — a new method for identification of **D**ifferentially **E**xpre**S**sed gene **MO**dules i**N** **D**iseases

① Assigning patients on edges

Aim: find connected pairs of genes up- or down-regulated in subgroups of patients compared to the background



- for every pair of connected genes, identify optimal pair of thresholds, such that
  - number of patients with expression above both thresholds is maximal
  - this patient overlap is significant (hypergeometric test)
- the same approach is used to identify pairs of down-regulated genes and corresponding patients

② Probabilistic edge clustering

Aim: find connected subnetworks with similar patient sets on edges

- X_{mxn} for n edges and m patients is a matrix of patient module memberships obtained in step 1, where $x_{ij} = 1$, if sample i is assigned to edge j, or 0 otherwise
- The assignments of samples to the edges are modeled as a Bernoulli distribution with parameter θ_{ic} for each samples i and module c:

$$x_{ji}|\theta_{ic}, s_j \sim Bernoulli(x_{ji}|\theta_{ts_j}), \quad \theta_{ic}|\alpha \sim Beta(\theta_{ic}|\alpha/2, \alpha/2) \; for \; 1 \le c \le K$$

- s_j indicates the module membership of edge j and follows a categorical distribution with parameter π and a Dirichlet prior:

$$s_j|\pi \sim Categorical(s_j|\pi), \quad \pi|\beta \sim Dirichlet(\underbrace{\beta/K, ..., \beta/K}_{K \text{ of them}})$$

- conditional probability of edge j to join the module k:

$$P(s_j = k|X, s_{-j}, \alpha, \beta) \propto \prod_{i:x_{ji}=1} [\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j} x_{li}}{\alpha + |\{l : s_l = k, l \neq j\}|}] \times \prod_{i:x_{ji}=0} [\frac{\alpha/2 + \sum_{l:s_l=k, l\neq j}(1 - x_{li})}{\alpha + |\{l : s_l = k, l \neq j\}|}] \times \frac{|\{l : s_l = k, l \neq j\}| + \beta/K}{m - 1 + \beta}$$

- Gibbs sampling is performed for parameter learning. After the model convergence, consensus module membership is taken.

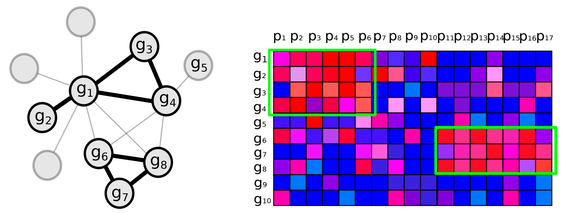
③ Postprocessing

- filtering out modules with low SNR
- merging modules with high overalp

PROBLEM DEFINITION

Find **groups of genes**

- **connected** on the PPI network
- **differentially expressed** in a **subgroup** of samples



Discovered **modules** are **reproducible**, enriched by **functionally related genes, and associated** with breast cancer **subtypes** and **survival**

This poster and code

ozolotareva@techfak.uni-bielefeld.de

https://github.com/ozolotareva/DESMOND

EVALUATION

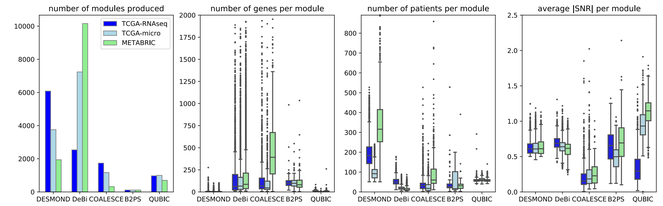
Datasets

- **TCGA-BRCA**
  - RNA-seq (1081)
  - microarrays (528)
- **METABRIC**
  - microarrays (1904)

Baseline Methods

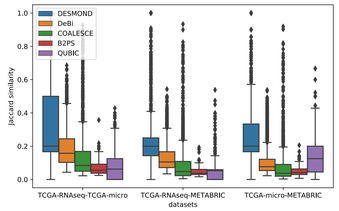
- DeBi
- QUBIC
- COALESCE
- B2PS

Characterization of modules produced by the five algorithms



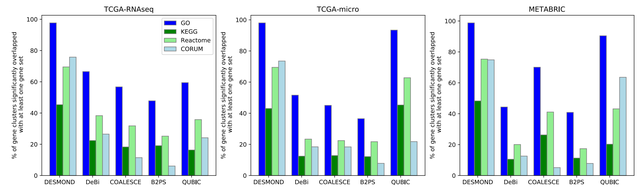
- DESMOND was the only method producing a number of large modules in terms of patients comprising nearly half of the whole cohort

Reproducibility



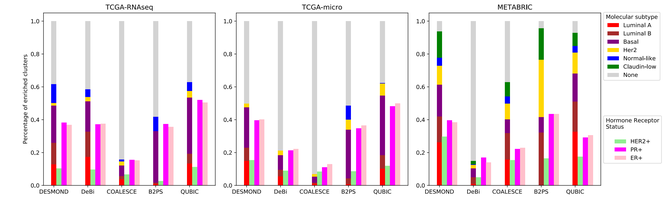
- Given clusterings $A=\{S_1,..,S_a\}$ and $B=\{S'_1,..,S'_b\}$, for every cluster $S_i \in A$ its best match $S'_j \in B$ is such that Jaccard similarity $J(S_i,S'_j)$ is maximal
- Similarly, all best matches of every $S'_j \in B$ are identified
- Distributions of Jaccard similarities of all matched pairs between the gene clusters found on datasets A and B are compared
- DESMOND produced more similar gene clusters in all three comparisons

GO and pathway enrichment



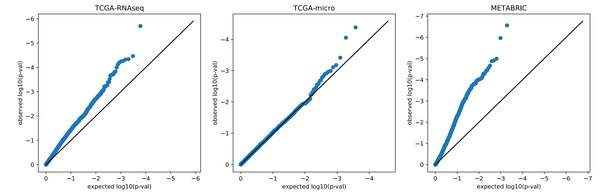
- DESMOND identified higher percentages of modules enriched by functionally related gene sets than the other methods

Association with known breast cancer subtypes



- DESMOND, QUBIC and B2PS identified similar proportions of sample clusters associated withknown molecular subtypes on all three datasets

Survival analysis



- Some modules identified by DESMOND are associated with overall survival