Involvement of RP11-334E6.12 lncRNA in gastric cancer pathogenesis

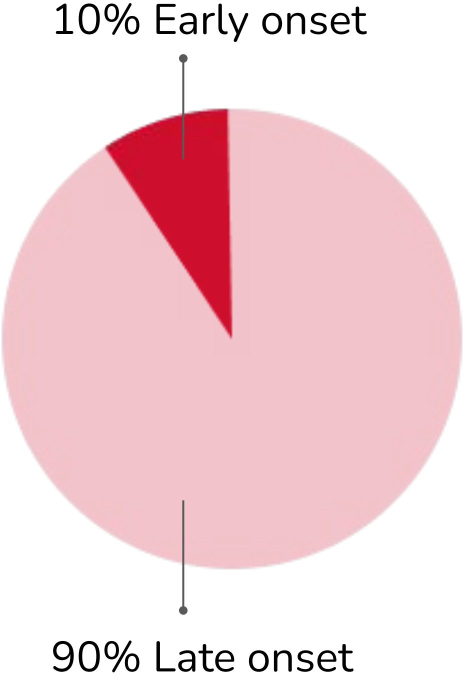
Stefano Cretti[1], Mario Lauria[2]

[1]Department of Cellular, Computational and Integrative Biology, University of Trento, Povo (TN), Italy
[2]The Microsoft Research, University of Trento Centre for Computational and Systems Biology, Rovereto (TN), Italy

Poster presentation for the “SIBBM 2022 • Frontiers in Molecular Biology”
conference (20-22 June 2022, Rome)
and the “Network based data analysis course”

Gastric cancer and study objective

- **5th most common** tumor overall
- **3rd cause of death** worldwide
- Before **45** years old -> **early onset gastric cancer** (EOGC), generally not due to carginogenic substances[1]
- **Current detection strategies**[2]:
  - invasive (**gastroscopy**)
  - aspecific (pepsinogen in the serum)

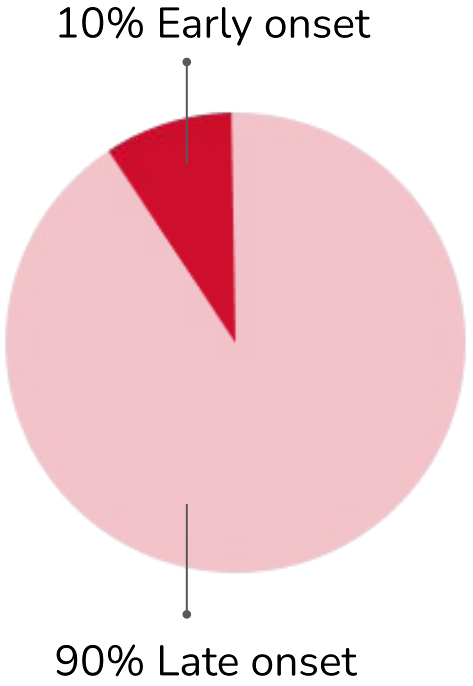


[1] Smyth et al., 2020, PMID: 32861308
[2] Necula et al., 2019, PMID: 31114131

Gastric cancer and study objective

- **5th most common** tumor overall
- **3rd cause of death** worldwide
- Before **45** years old -> **early onset gastric cancer** (EOGC), generally not due to carginogenic substances[1]
- **Current detection strategies**[2]:
  - invasive (**gastroscopy**)
  - aspecific (pepsinogen in the serum)

Identify candidate molecules in order to study their **serum or EVs concentration** and their potential applications as **bio-markers** for easy **screening**, early **detection**, post-surgery **follow-up.**

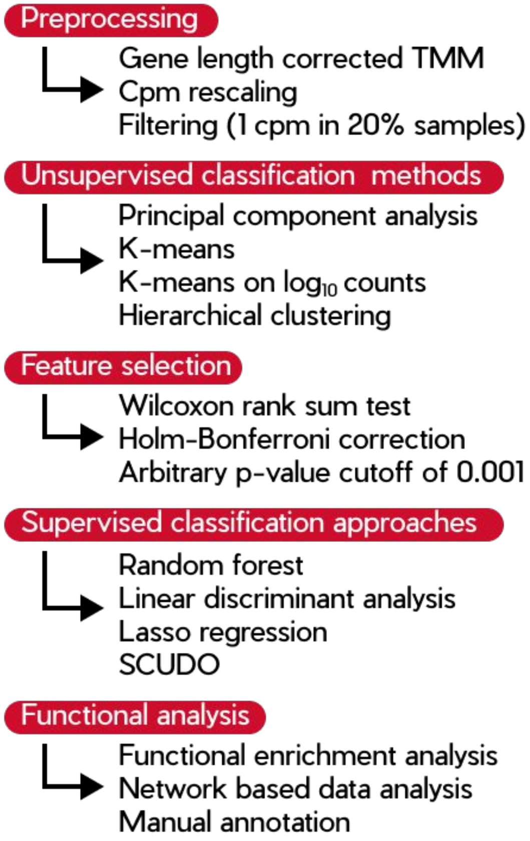


[1] Smyth et al., 2020, PMID: 32861308
[2] Necula et al., 2019, PMID: 31114131

Data and methods

Dataset publicly available on recount3, (SRA Study: SRP172499)[3]

RNA-Seq data of paired healthy stomach and GC samples for 79 EOGC patients (mostly diffuse-type) plus 1 GC sample from another EOGC patient



Preprocessing

- Gene length corrected TMM
- Cpm rescaling
- Filtering (1 cpm in 20% samples)

Unsupervised classification methods

- Principal component analysis
- K-means
- K-means on $\log_{10}$ counts
- Hierarchical clustering

Feature selection

- Wilcoxon rank sum test
- Holm-Bonferroni correction
- Arbitrary p-value cutoff of 0.001

Supervised classification approaches

- Random forest
- Linear discriminant analysis
- Lasso regression
- SCUDO

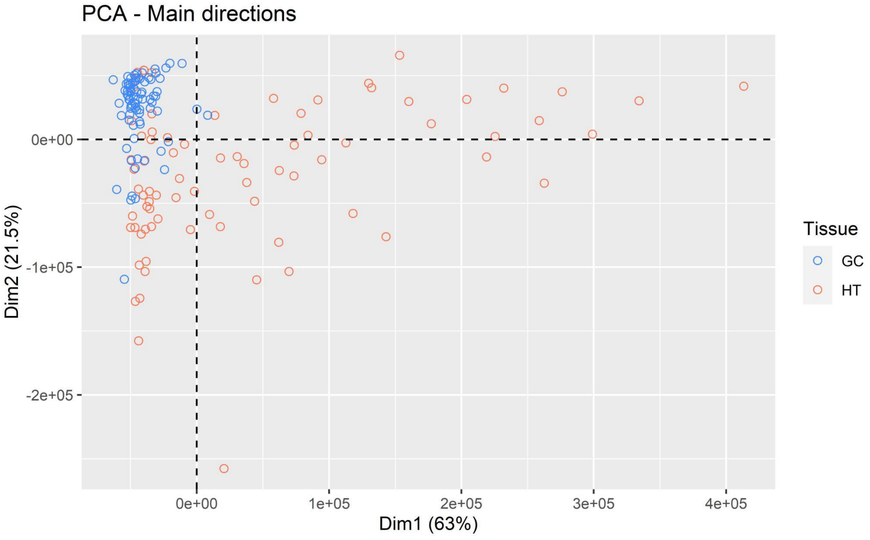
Functional analysis

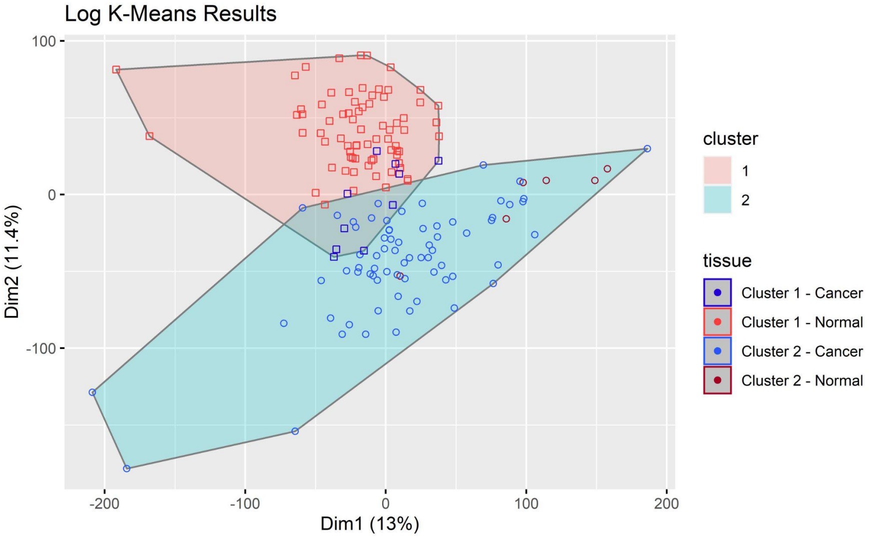
- Functional enrichment analysis
- Network based data analysis
- Manual annotation

[3] Mun et al., 2019, PMID: 30645970

Unsupervised methods

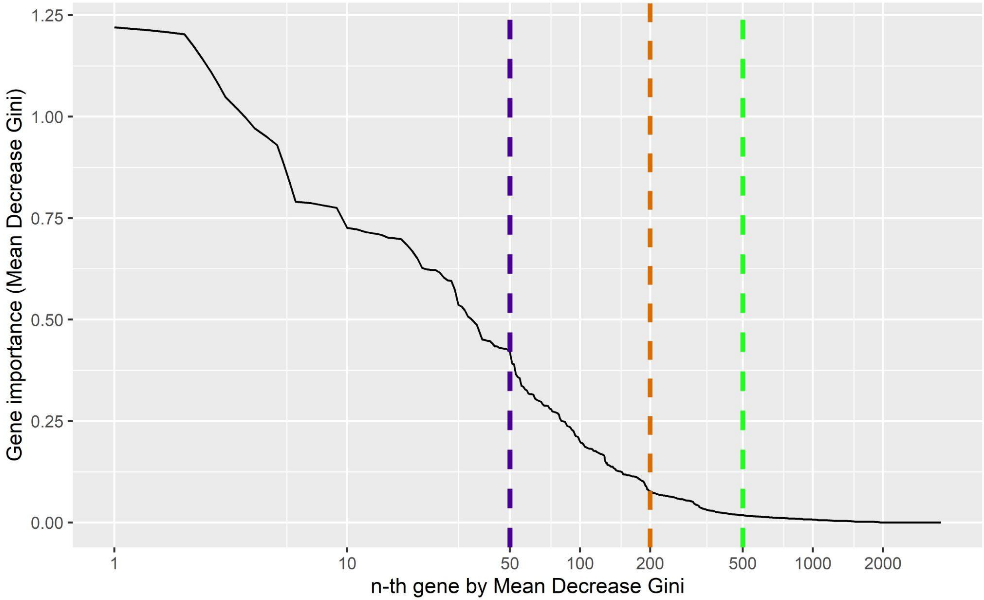
- **PCA**: tight cluster of cancer samples, spread of healthy samples
- **K-means**: poor clustering due to high spread
- **Log K-means**: better clustering, could maybe be optimized
- **Hierarchical clustering**: many single branches, poor division





Supervised methods

- **Random forest**: very good classification
- **LDA**: good results but not as good as random forest
- **Lasso regression**: good classification but using aspecific genes
- **SCUDO**: good results but not as good as random forest

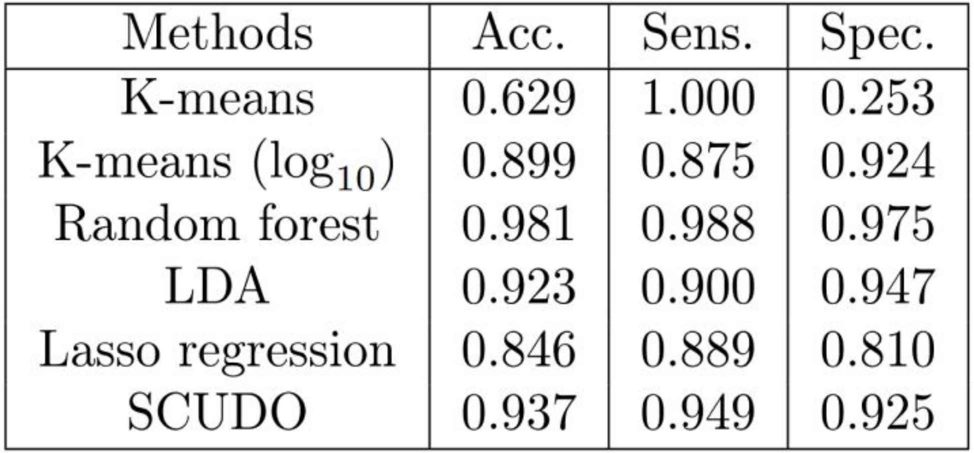


Classifier comparison

Most methods perform well

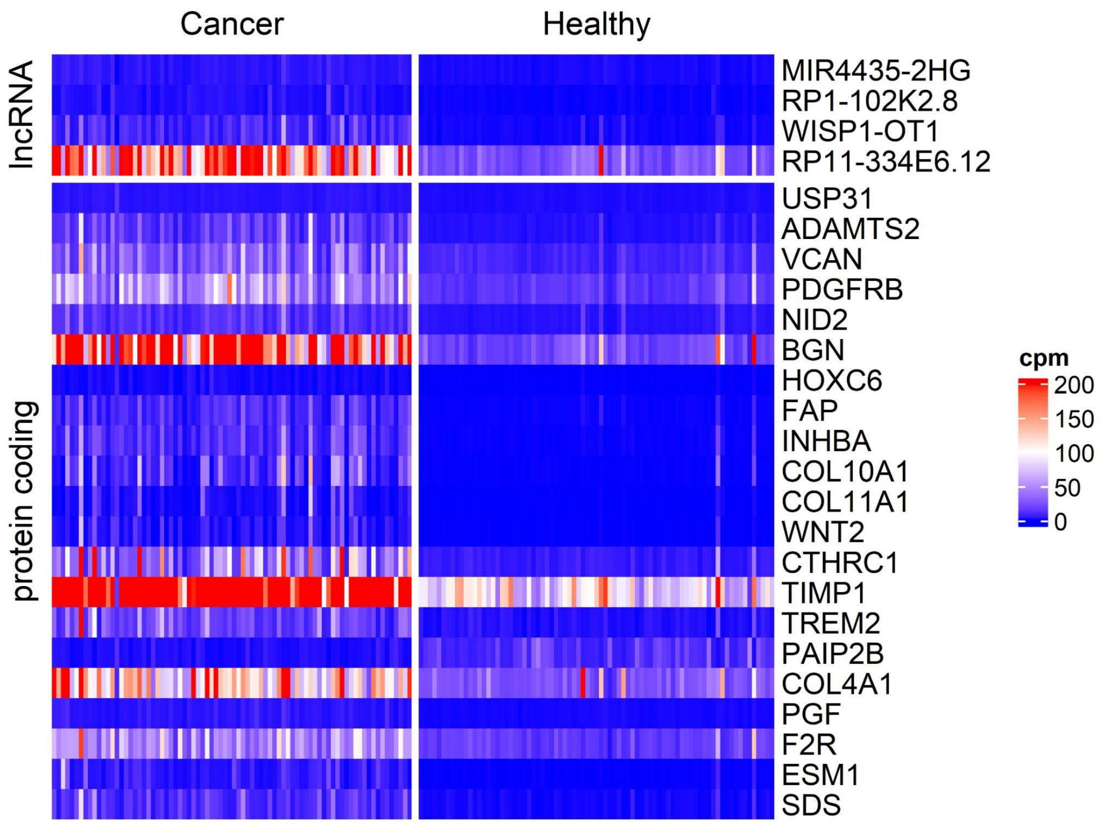
Possible further optimization for potentially better results

Random forest after feature selection is the best performing classifier



Methods	Acc.	Sens.	Spec.
K-means	0.629	1.000	0.253
K-means ($\log_{10}$)	0.899	0.875	0.924
Random forest	0.981	0.988	0.975
LDA	0.923	0.900	0.947
Lasso regression	0.846	0.889	0.810
SCUDO	0.937	0.949	0.925

Most influential genes

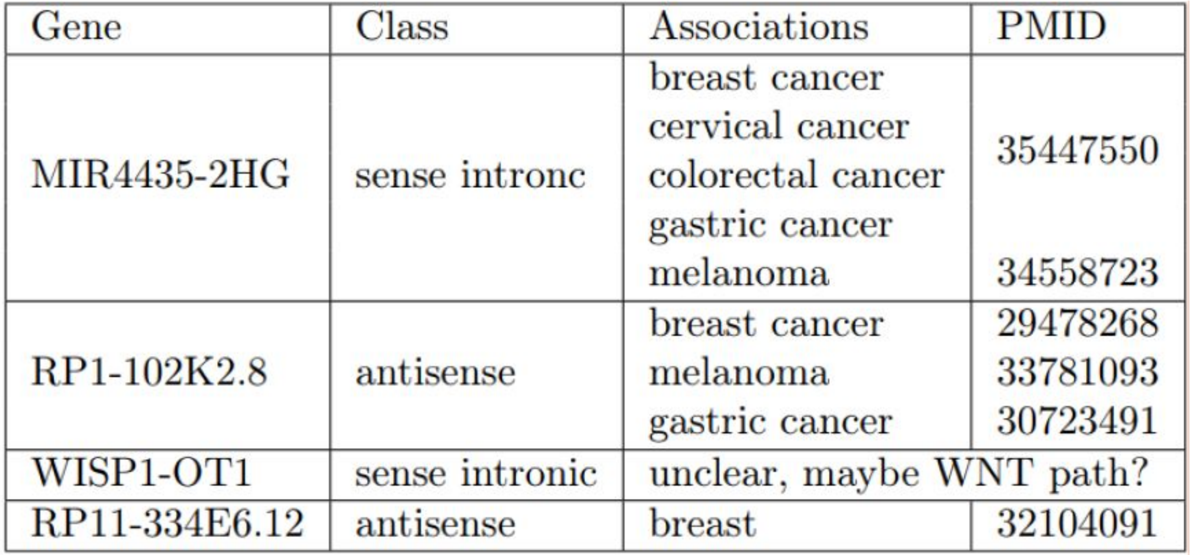


Top 200 genes by contribution to random forest classifier were selected for downstream analysis

Top 25 genes are displayed in the heat map (cpm values above 200 are capped)

4 of the top 25 genes are lncRNA, all of which more expressed in GC than in HT

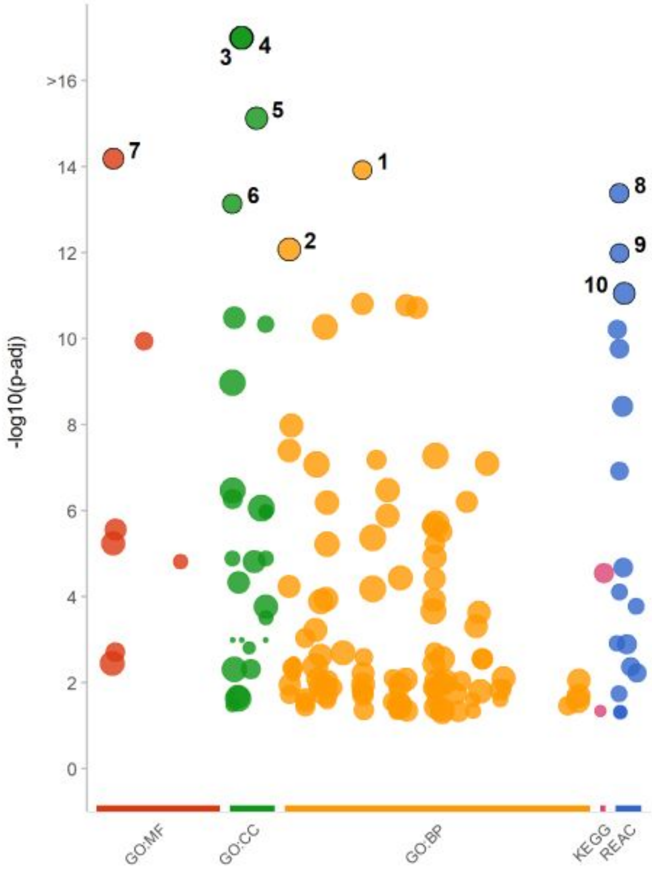
Most influential lncRNA



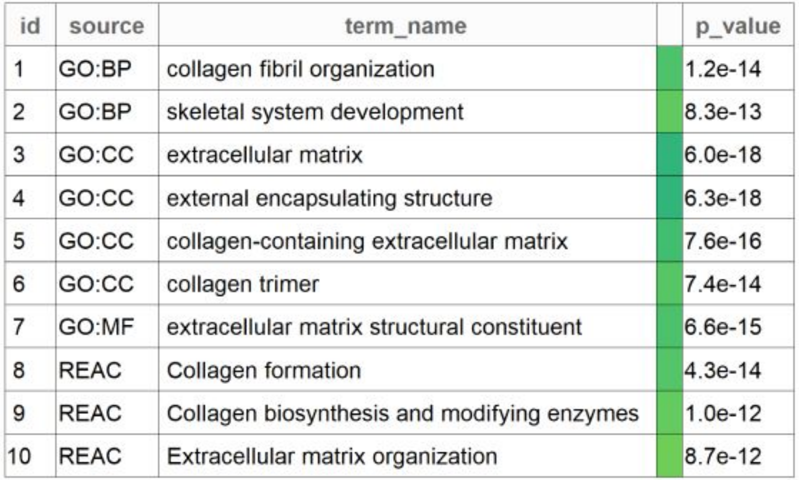
Gene	Class	Associations	PMID
MIR4435-2HG	sense intronc	breast cancer cervical cancer colorectal cancer gastric cancer melanoma	35447550 34558723
RP1-102K2.8	antisense	breast cancer melanoma gastric cancer	29478268 33781093 30723491
WISP1-OT1	sense intronic	unclear, maybe WNT path?	
RP11-334E6.12	antisense	breast	32104091

RP11-334E6.12 lncRNA particularly interesting, with no prior connection to gastric cancer and with **THY1** (CD90) as putative target (potential role in downregulating immune response?)

Gene set enrichment analysis

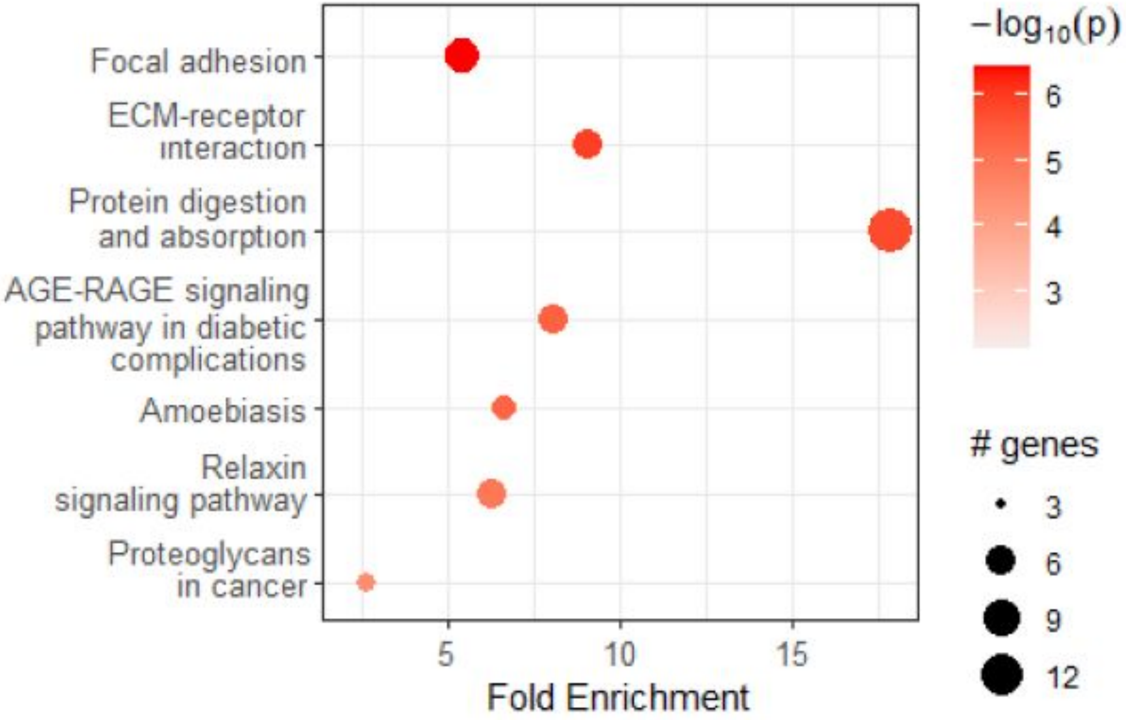


Gene enrichment analysis displays an overrepresentation of **extracellular matrix** and **tissue remodelling genes**



id	source	term_name		p_value
1	GO:BP	collagen fibril organization	🟩	1.2e-14
2	GO:BP	skeletal system development	🟩	8.3e-13
3	GO:CC	extracellular matrix	🟩	6.0e-18
4	GO:CC	external encapsulating structure	🟩	6.3e-18
5	GO:CC	collagen-containing extracellular matrix	🟩	7.6e-16
6	GO:CC	collagen trimer	🟩	7.4e-14
7	GO:MF	extracellular matrix structural constituent	🟩	6.6e-15
8	REAC	Collagen formation	🟩	4.3e-14
9	REAC	Collagen biosynthesis and modifying enzymes	🟩	1.0e-12
10	REAC	Extracellular matrix organization	🟩	8.7e-12

Network based data analysis



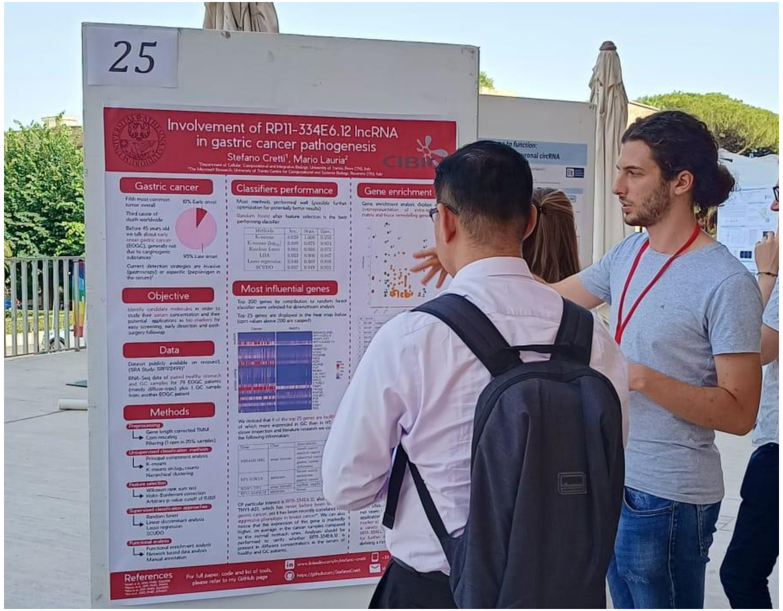
Network analysis produces mostly similar results to gene enrichment, meaning **ECM linked genes overrepresentation**, but it also highlights **tissue specific genes**

Conclusions

Discerning healthy and tumoral tissue from RNA-seq of bioptic material does not seem to be too challenging; yet this application is fairly limited and a serum marker could prove extremely valuable for a variety of reasons.

We thus **propose RP11-334E6.12 lncRNA as candidate target** for further analyses aimed at potentially defining a **novel bio-marker**.

Contacts



The poster for this project won the "Riccardo Cortese Award" for Best Poster at SIBBM 2022 - Frontiers in Molecular Biology (Rome 20-22 June)

For full paper, code and list of tools, please refer to my GitHub page

www.linkedin.com/in/stefano-cretti

☎ +39 3407487765

https://github.com/StefanoCretti

✉ stefano.cretti@studenti.unitn.it