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).

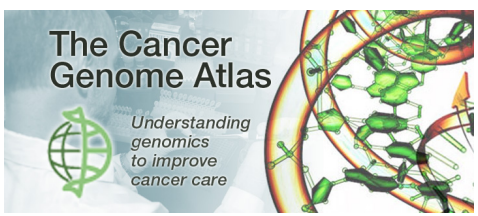
Extend these methods to a **pancancer** analysis.

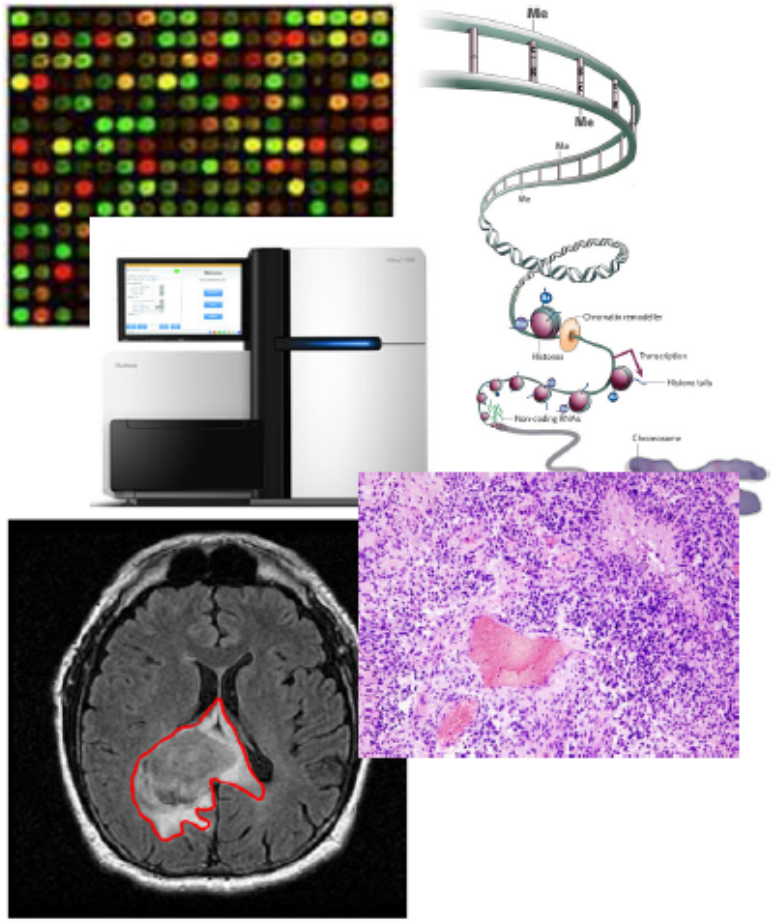
. . . . . Face the **big data** challenge . . . . . .

Data overview

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

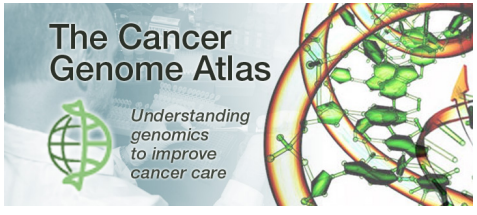
- Gene & miRNA expression (Agilent & Affy microarray - RNA sequencing)
- Copy number (Affy SNP 6.0)
- DNA methylation (Agilent Infinium (27k))
- Mutation (DNA sequencing)
- Pathology images
- Medical images (MRI, CT)





Data overview

Cancer Type	TCGA code	Samples	Genes
Bladder cancer	BLCA	181	15,432
Breast cancer	BRCA	985	16,020
Colorectal cancer	COADREAD	589	15,533
Glioblastoma	GBM	501	17,811
Head and Neck squamous carcinoma	HNSC	371	15,828
Kidney clear cell carcinoma	KIRC	509	16,123
Acute myeloid carcinoma	LAML	173	14,296
Lung adenocarcinoma	LUAD	489	16,092
Lung squamous carcinoma	LUSC	490	16,219
Ovarian cancer	OV	541	17,814
Endometrial cancer	UCEC	508	15,706



AMARETTO :

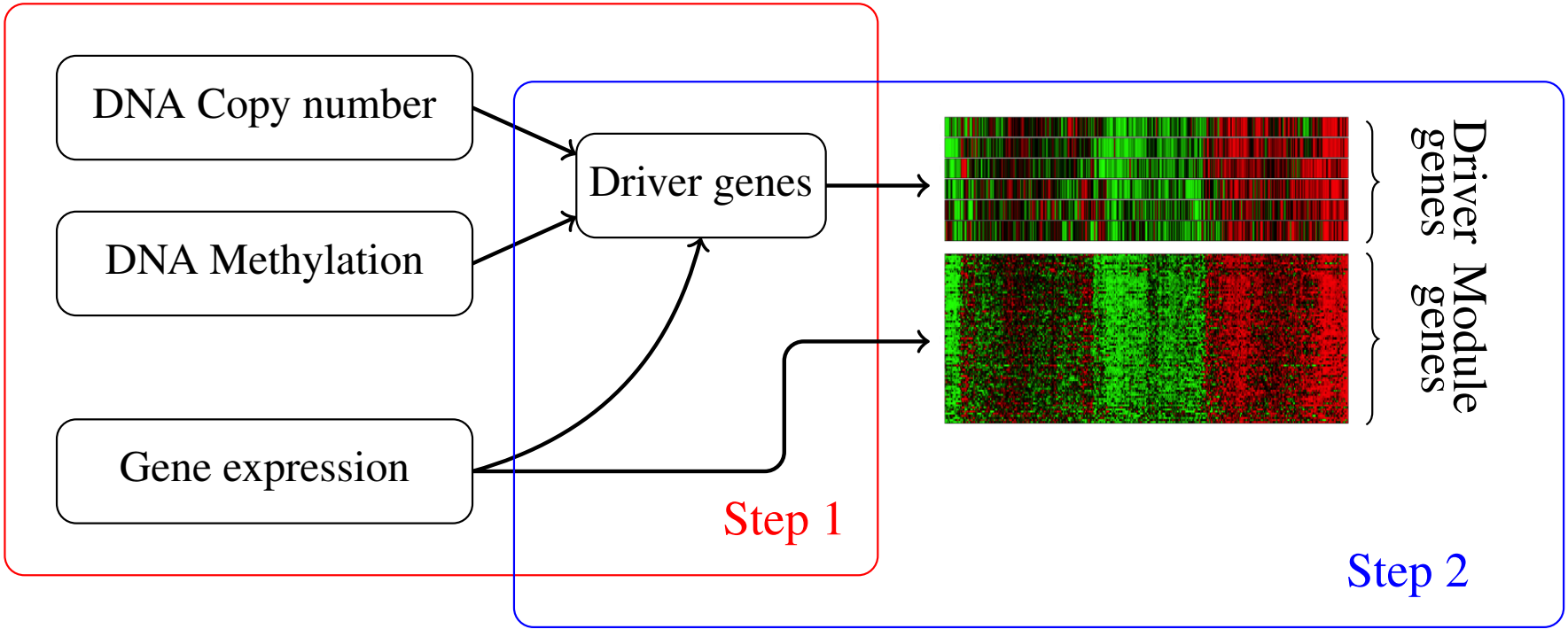
Multi-omics data fusion for cancer data

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.



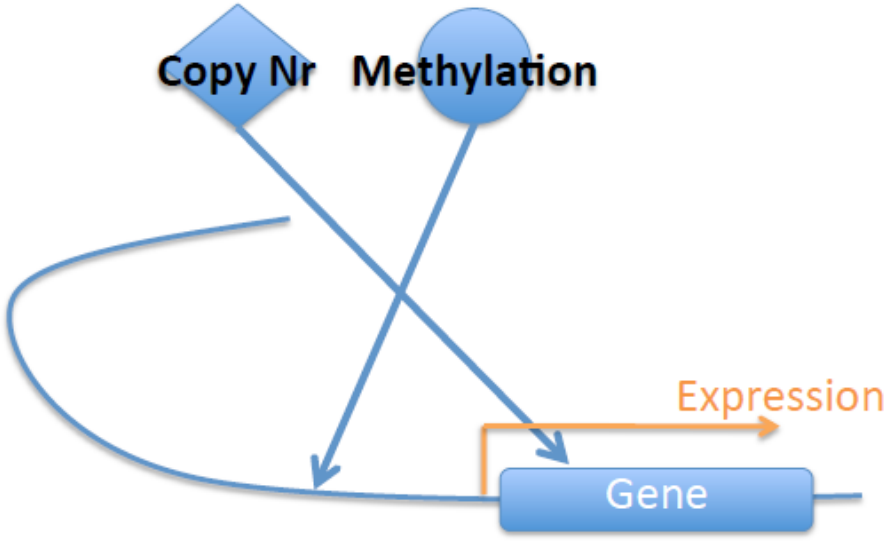
Gevaert et al., Interface Focus, 2013

Step 1 : Generating the list of candidate drivers

If gene expression can be explained by genomic events

↓

Candidate driver gene



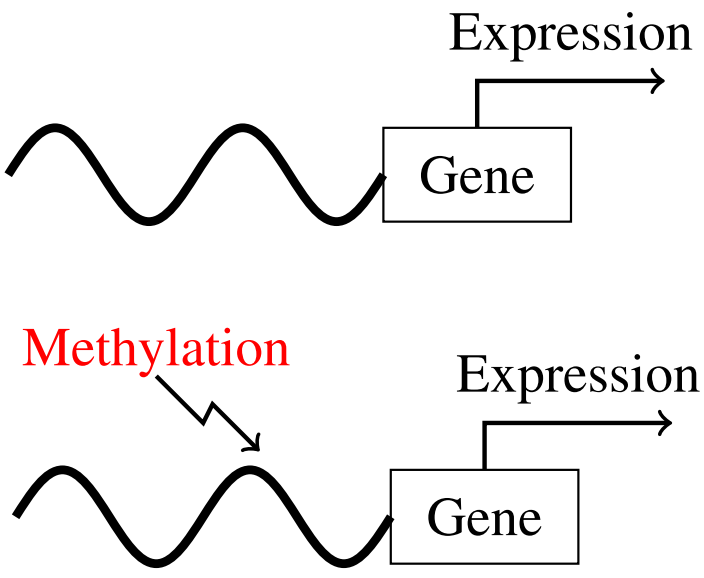
Model gene expression as a function of copy number and DNA methylation :

$$\text{Expression}_{\text{Gene}_i} = f(\beta_1 \underbrace{\text{Methylation}_{\text{Gene}_i}}_{\textit{MethylMix}} + \beta_2 \underbrace{\text{Copy Number}_{\text{Gene}_i}}_{\textit{GISTIC}}).$$

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)

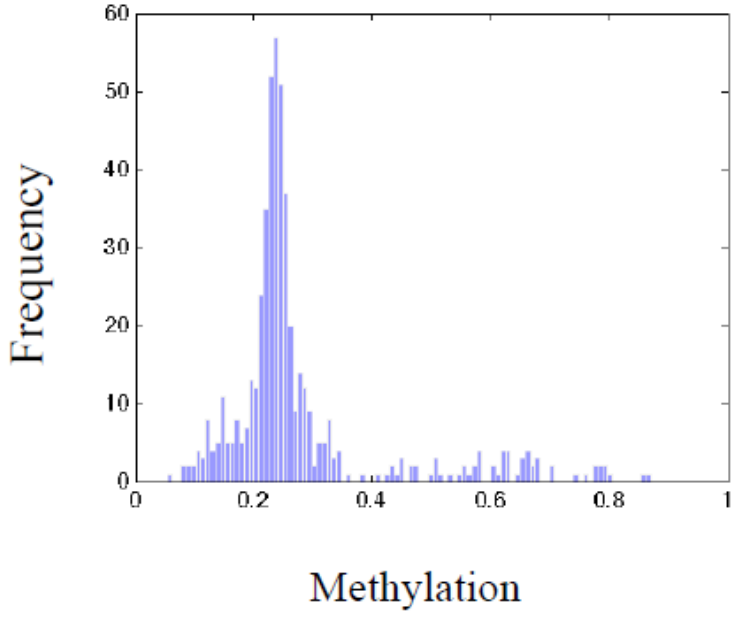
Step 1 : Cancer driver gene filtering (MethylMix)

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

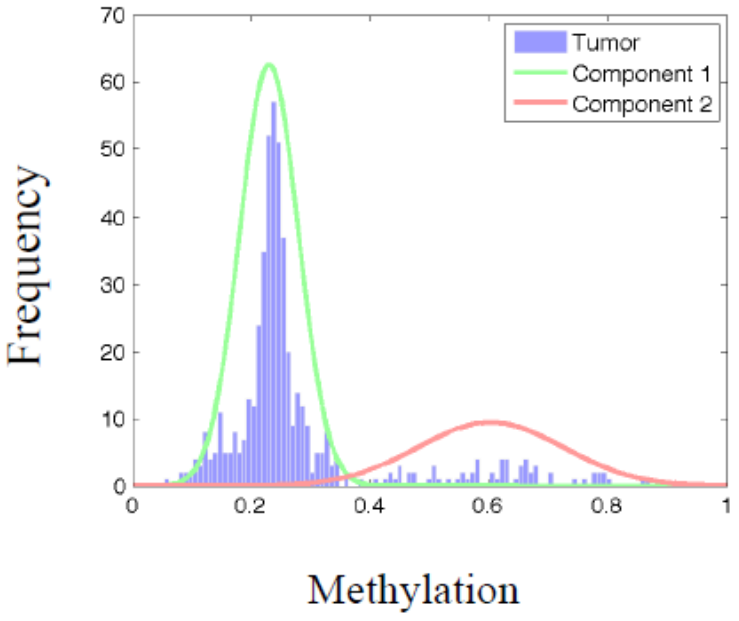


Gevaert et al., Genome Biology, 2015

Step 1 : Cancer driver gene filtering (MethylMix)

Step 2 :

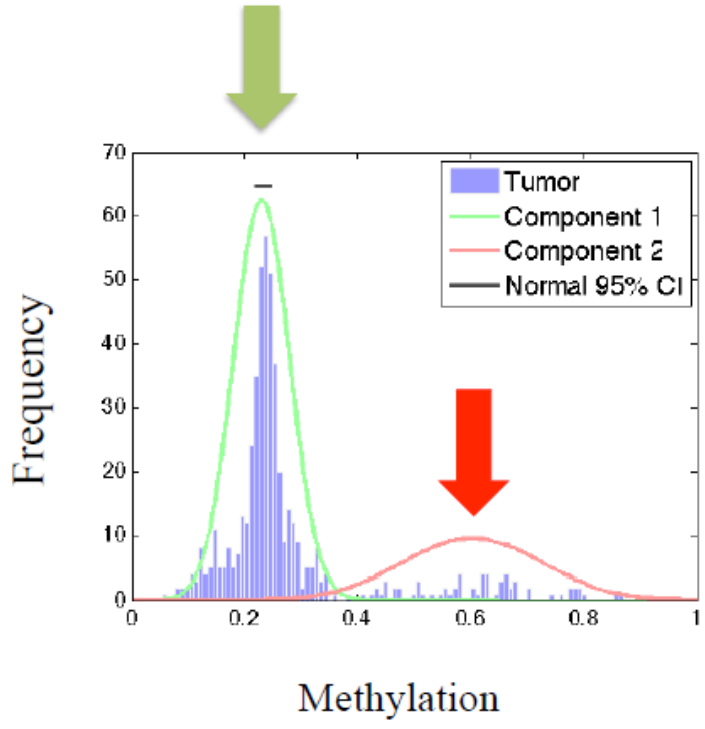
- mixture of beta distributions
- identification of two components



Step 1 : Cancer driver gene filtering (MethylMix)

Step 3 :

- comparison with DNA methylation in normal samples



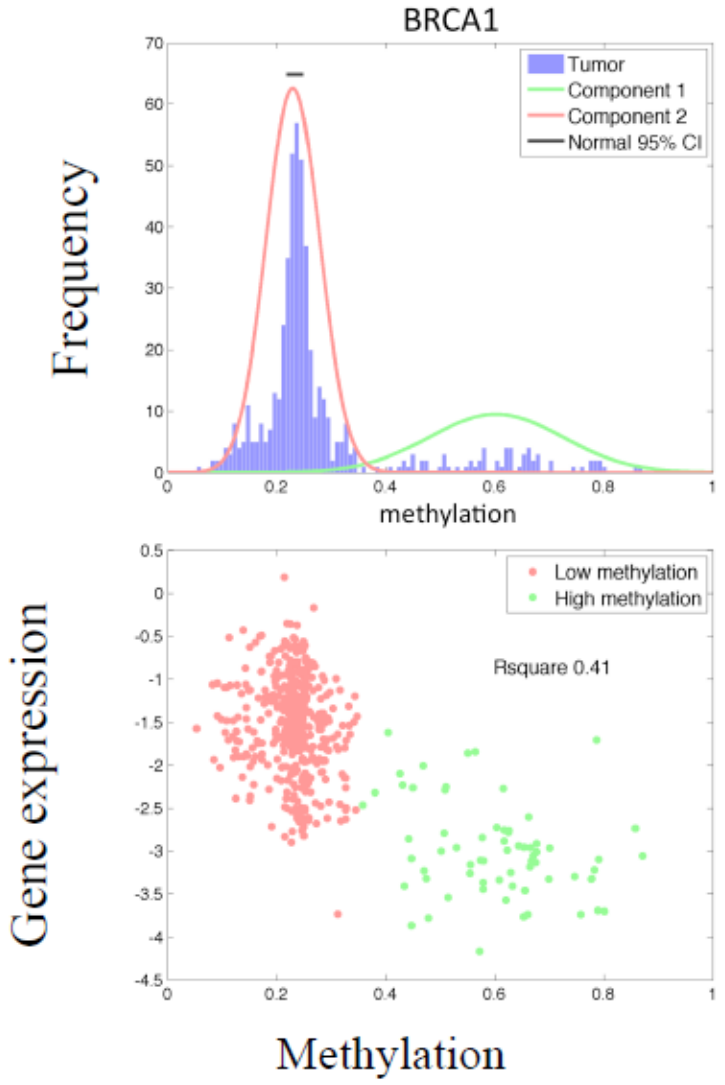
An example of hyper-methylation of BRCA1 in ovarian cancer

Step 1 : Cancer driver gene filtering (MethylMix)

Step 4 :

- inverse correlation with gene expression

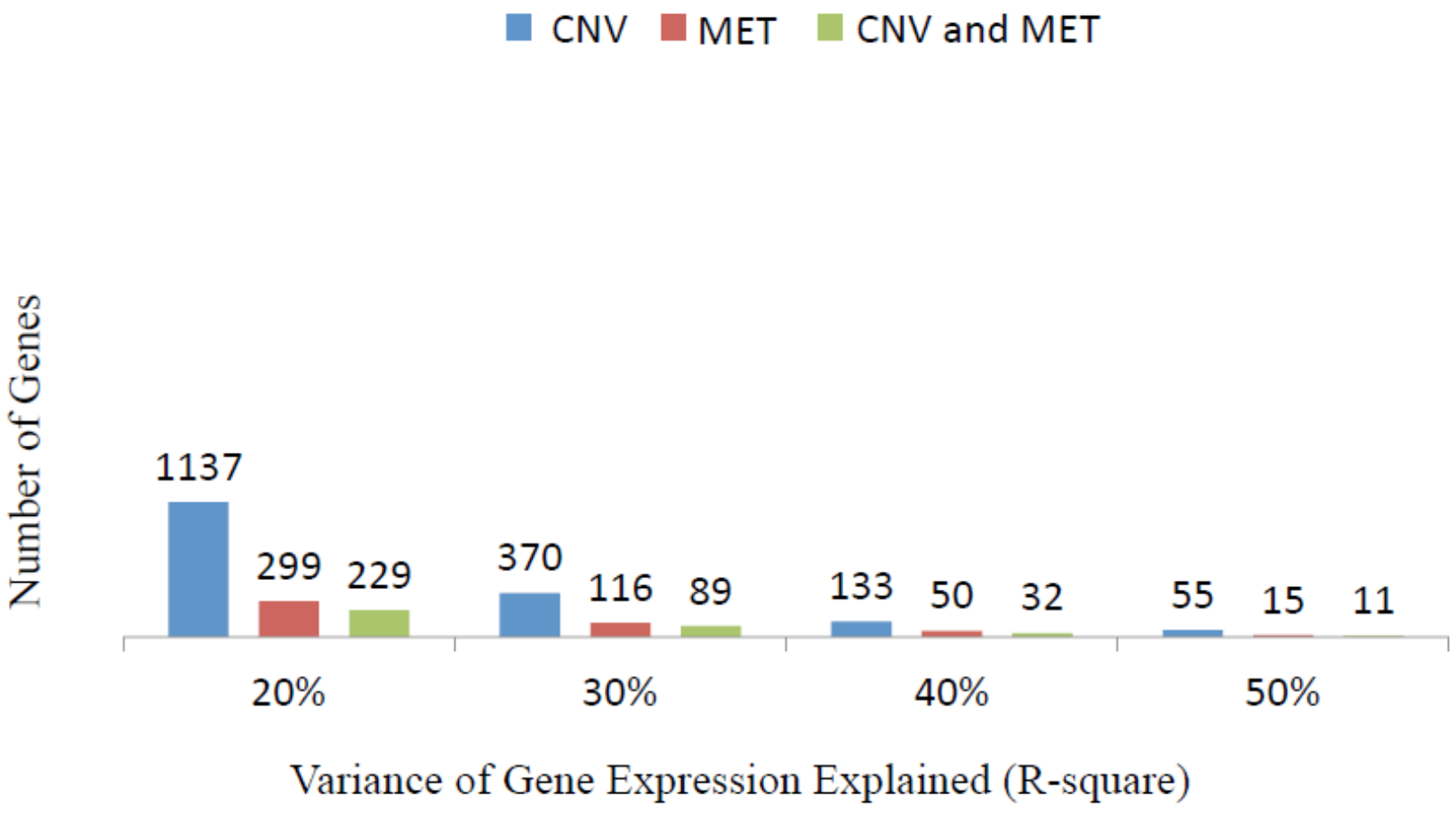
R-square statistic to quantify amount of variation explained



MethylMix results

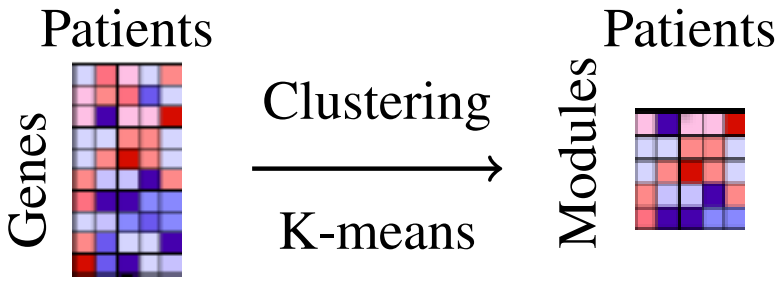
TCGA code	Total of Genes	Hyper-methylated	Hypo-methylated
BLCA	15,432	443	74
BRCA	16,020	798	203
COADREAD	15,533	847	177
GBM	17,811	246	140
HNSC	15,828	728	101
KIRC	16,123	319	251
LAML	14,296	470	77
LUAD	16,092	576	182
LUSC	16,219	605	133
OV	17,814	234	229
UCEC	15,706	618	238

Step 1 : Results for glioblastoma

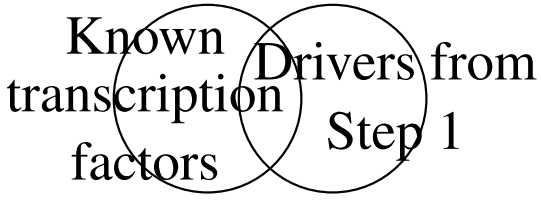


Step 2 : Associating candidate drivers with their downstream targets

Modules of co-expressed genes



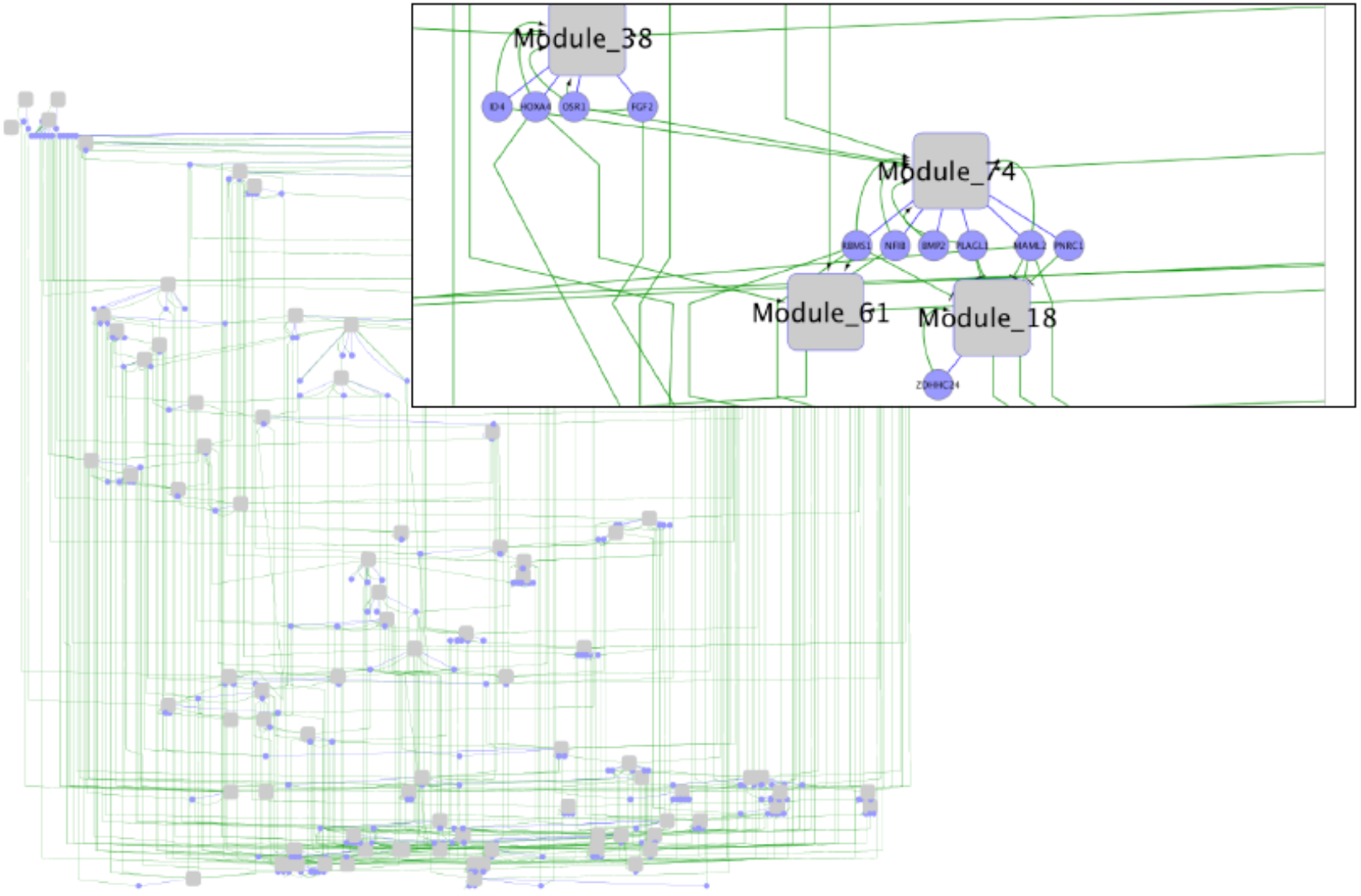
Set of driver genes



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

Linear regression + lasso regularization

Module network



AMARETTO captures drivers of cancer

Results obtained for **Glioblastoma**

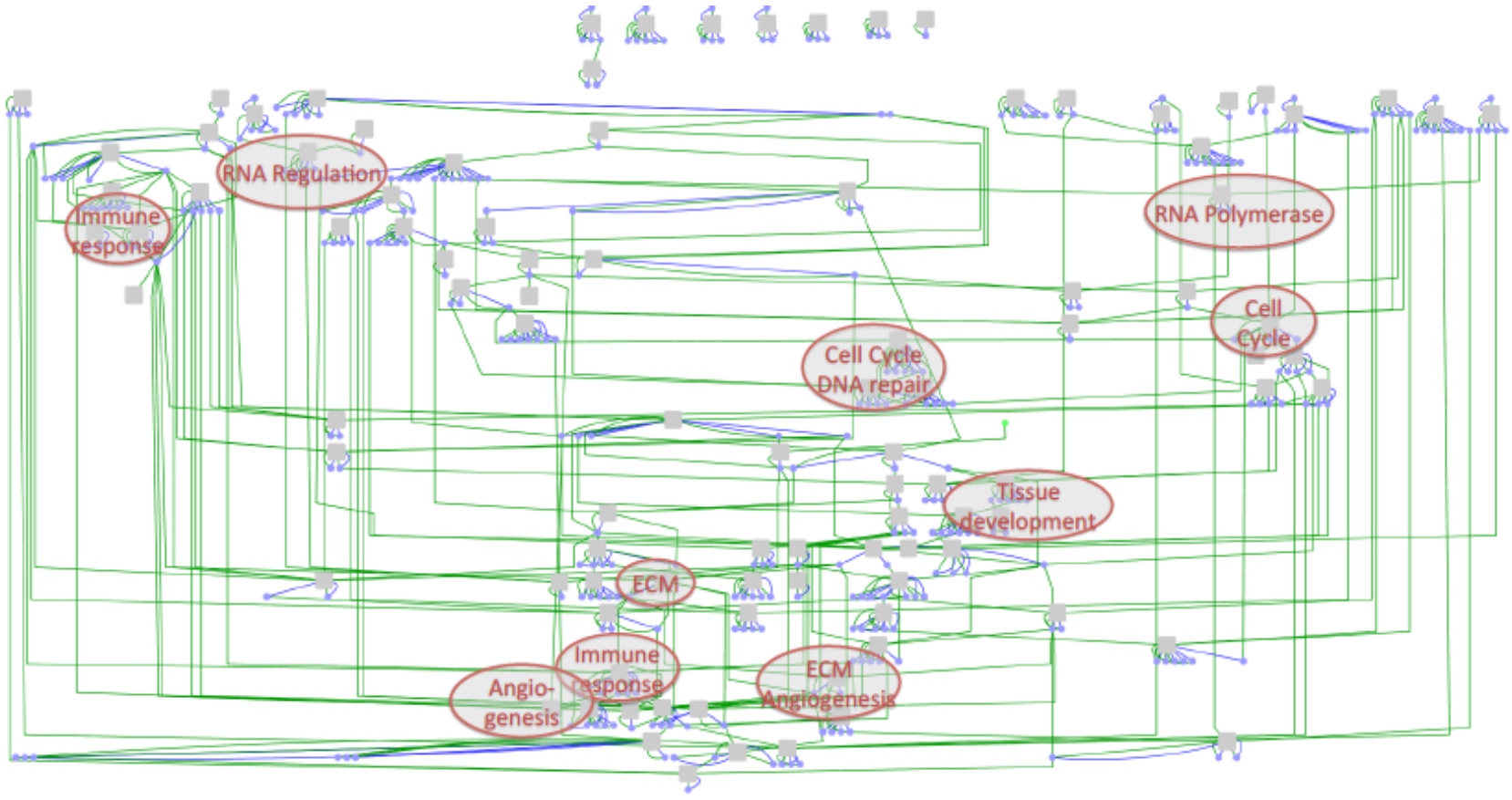
Cancer driver	Number of modules
ZNF300	10
TNFRSF1A	10
PTRF	8
WWTR1	8
MYT1	7
PYCARD	7
PATZ1	7
BASP1	6
RAB32	6
SATB1	6

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

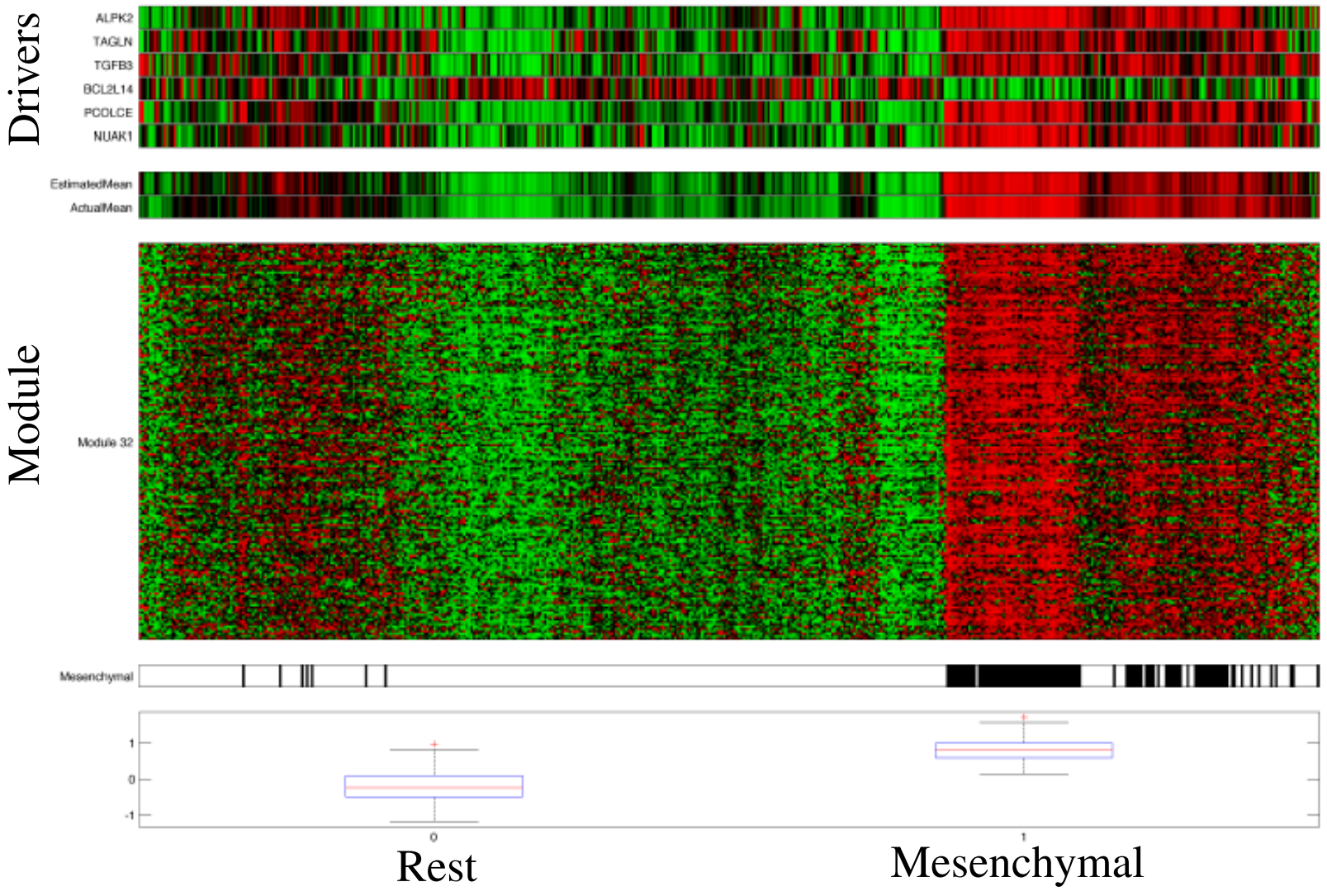
TCGA molecular subtypes of ovarian cancer :

- immuno-reactive
- differentiated
- mesenchymal
- proliferative

$\rightarrow$ modules **correlated** with subtypes point to potential driver genes

Bell et al., Nature, 2011

AMARETTO identifies drivers of subtypes



To a pancancer AMARETTO analysis ?

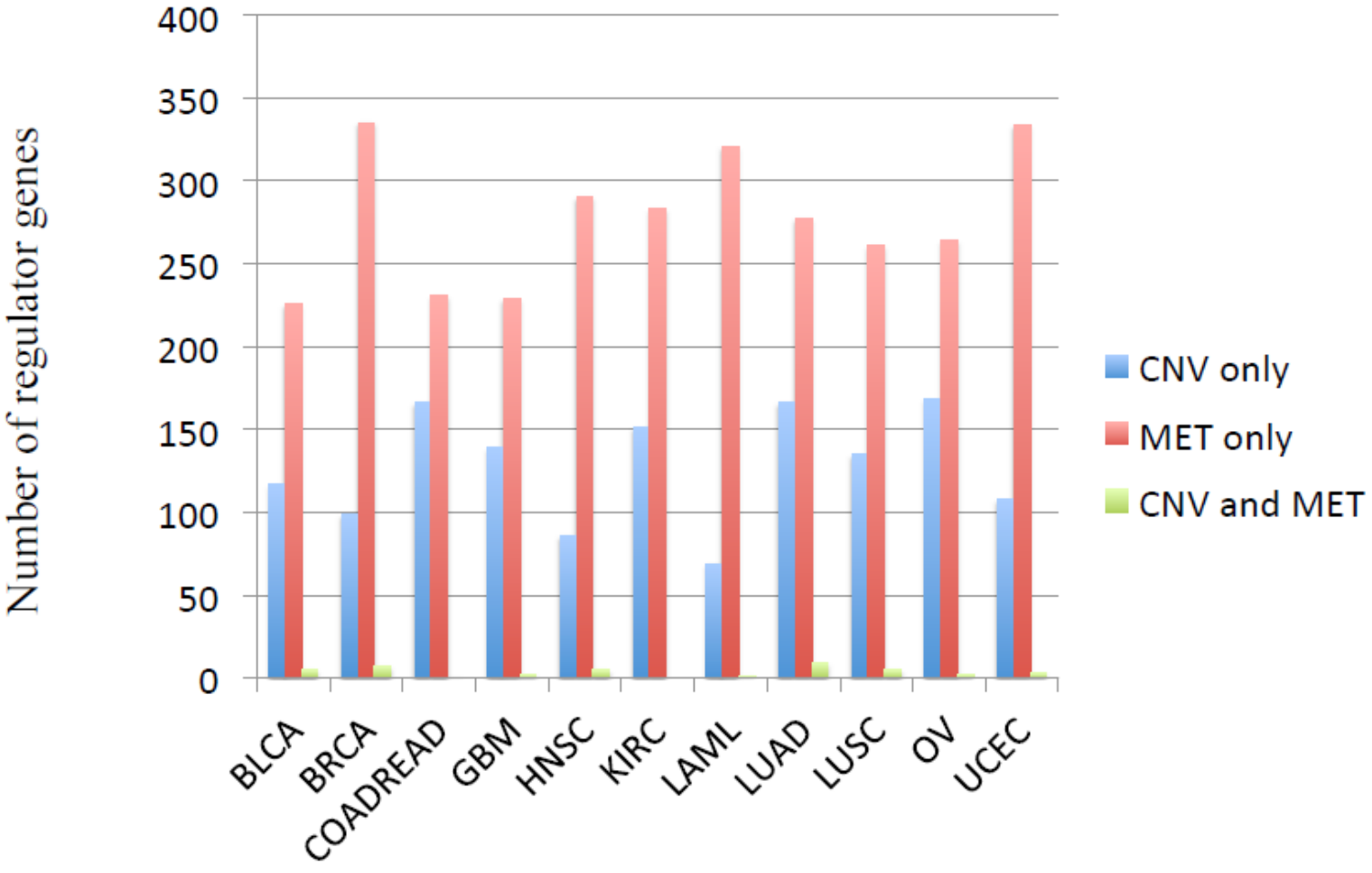
After running AMARETTO on the 11 cancer sites, 11 module networks were produced, with :

- 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,
  $\rightarrow$ sparse method
  $\rightarrow$ most of them are methylated
- 45 drivers regulate more than 15 modules across all cancer sites.

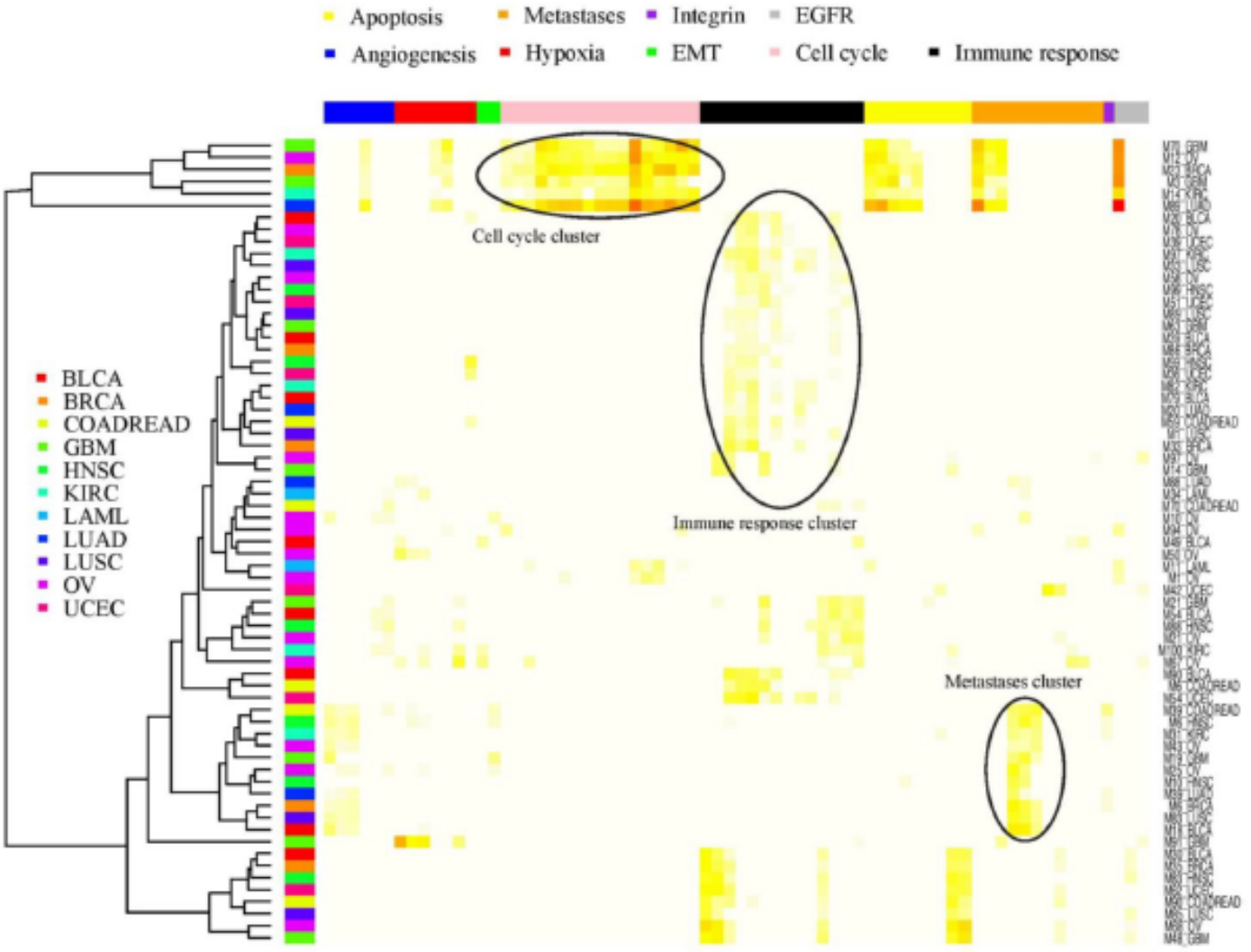
To a pancancer AMARETTO analysis ?



To a pancancer AMARETTO analysis ?

	Number of regulated modules	Number of involved cancers
FSTL1	31	7
IFFO1	29	6
MLPH	28	5
SPARCL1	26	5
CLIP3	24	4
MFAP4	24	4
BEND5	24	4
NUAK1	23	6
CAPS	23	9
PP1R16B	23	5
OLFML1	23	5
SLA	20	3
DDR2	20	7

To a pancancer AMARETTO analysis ?



AMARETTO :

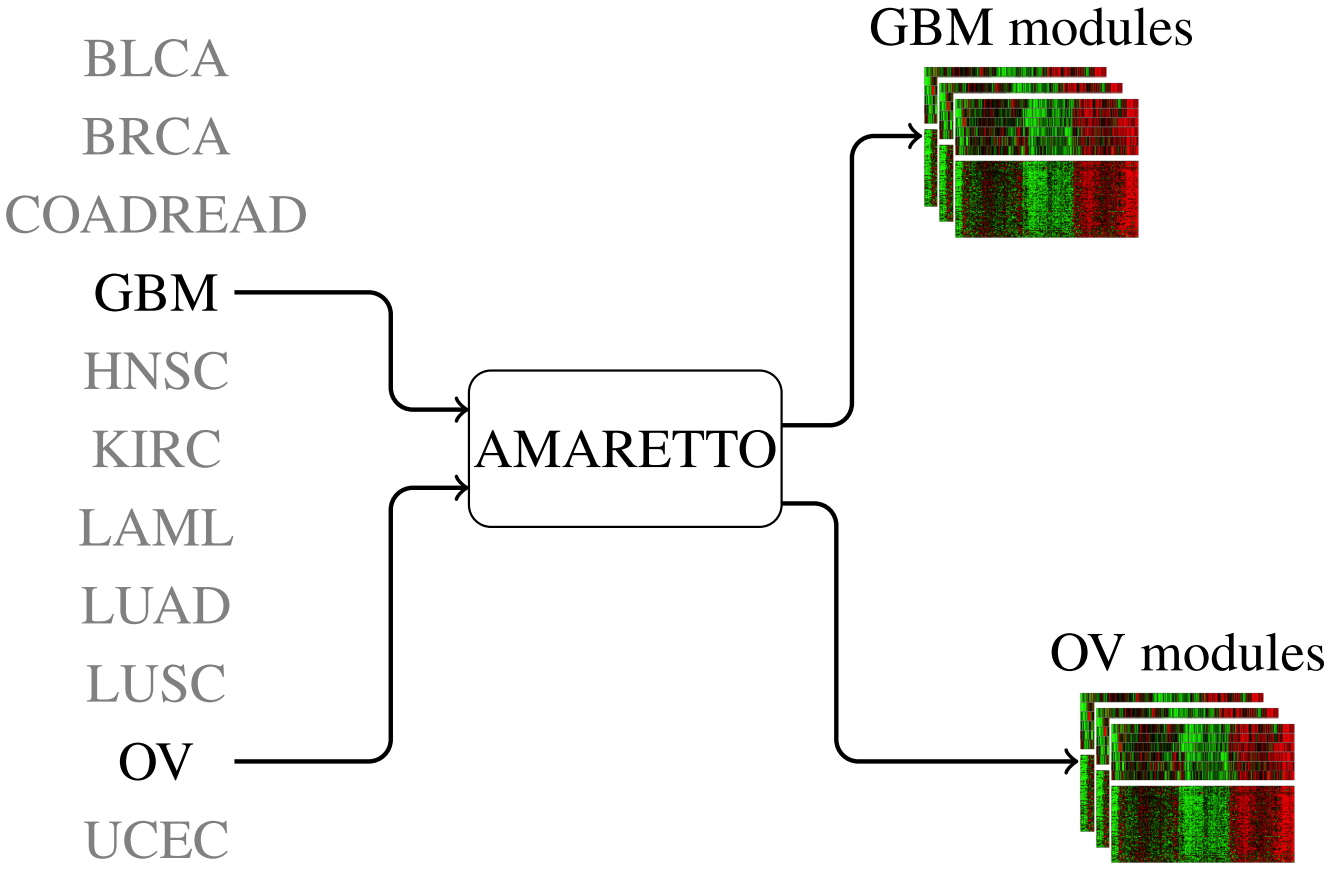
Multi-omics data fusion for cancer data

Pancancer module networks

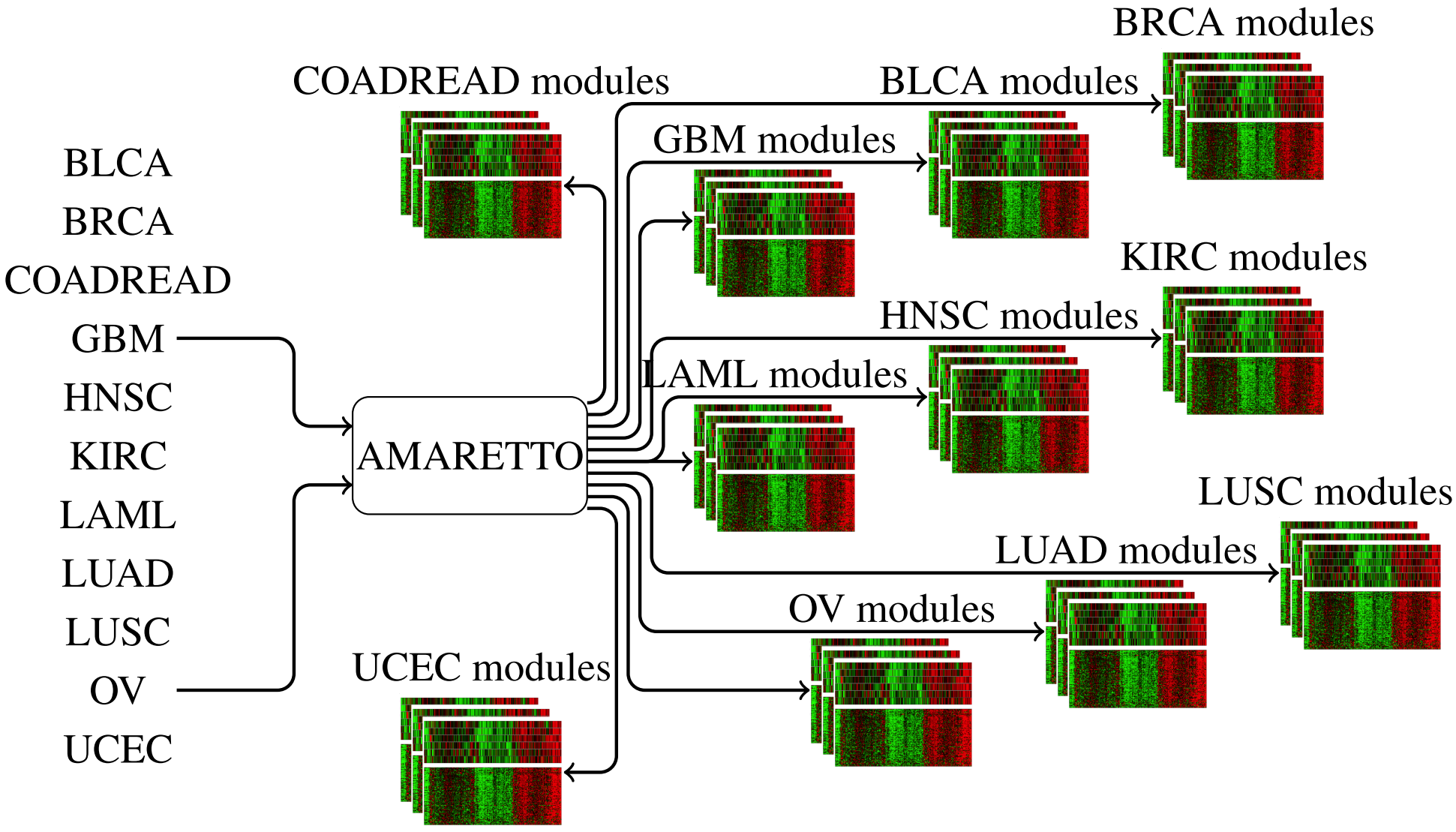
UNIVERSITÉ PARIS DESCARTES

Stanford MEDICINE

Pancancer analysis

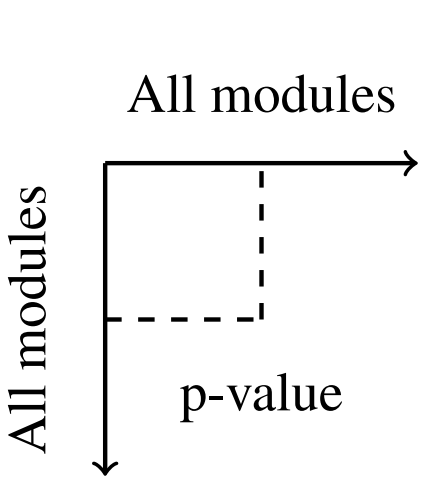


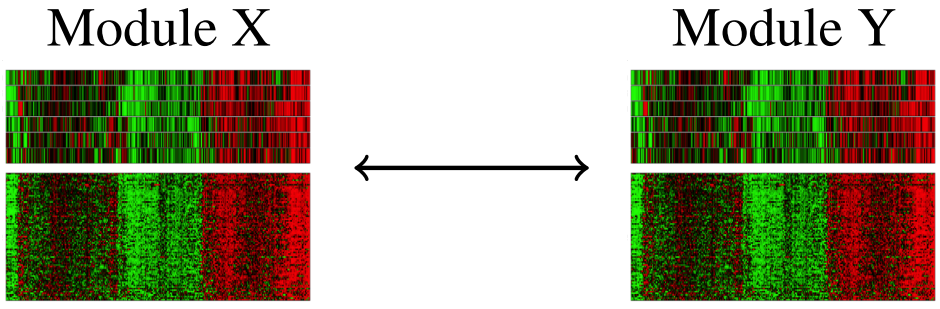
Pancancer analysis



Pancancer module network

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

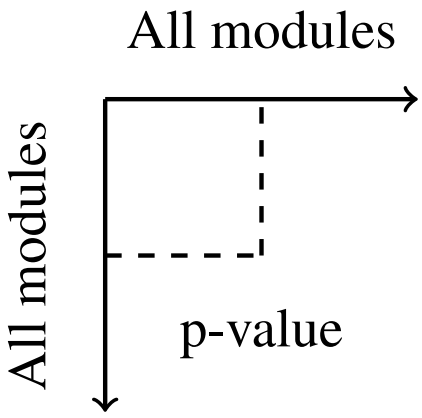


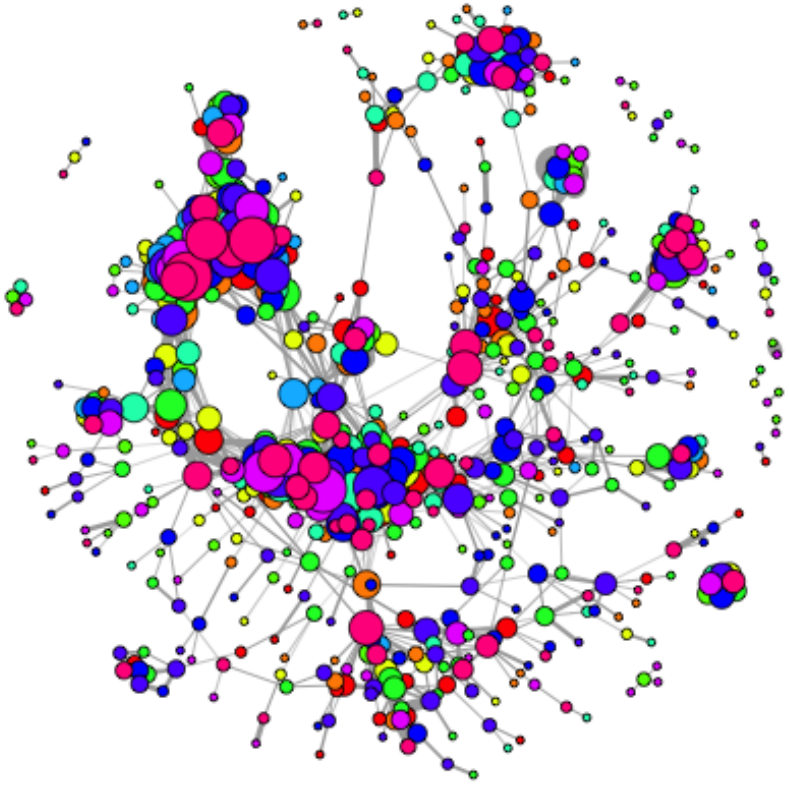


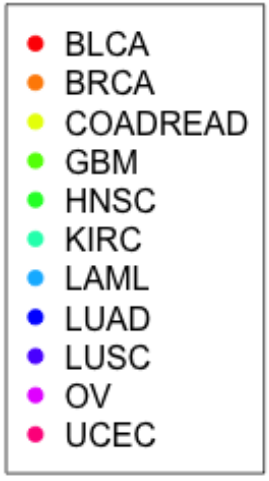
- p-value to compare modules
- enrichment score using negative log of the p-value

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 betweeness detection algorithm), which consists in :

1- computing the betweeness 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.

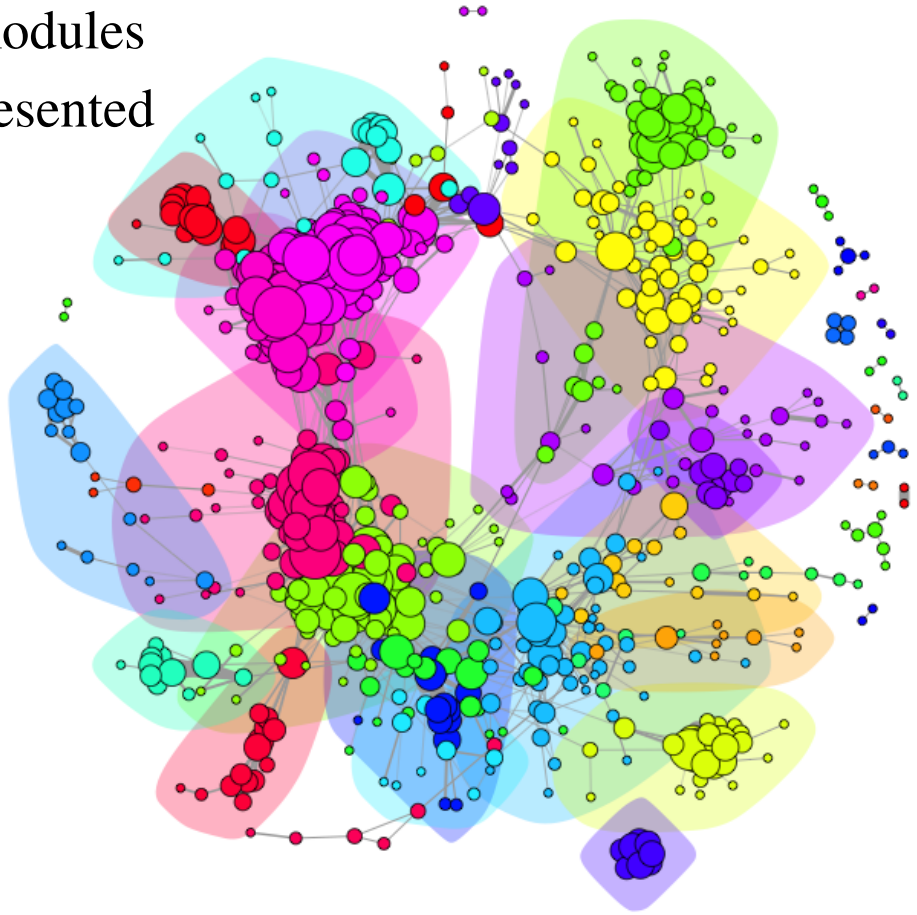
Weight edge betweenness score with the $-\log$ p-value score.

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

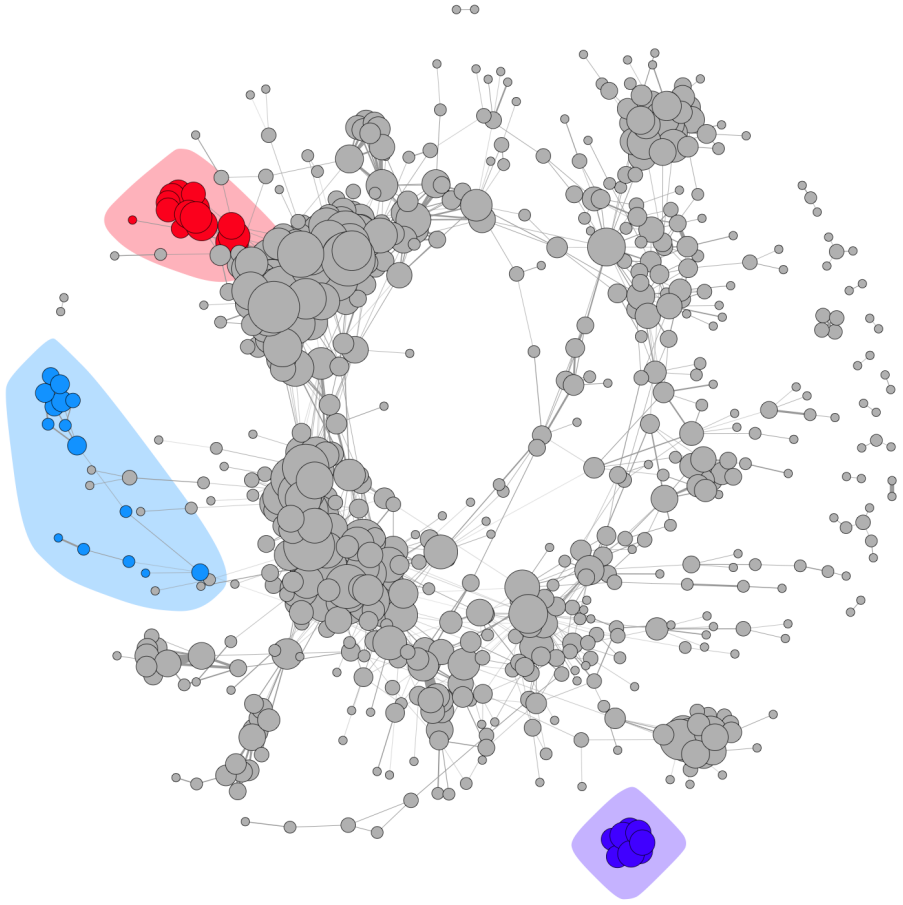


Pancancer histone community

· 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)

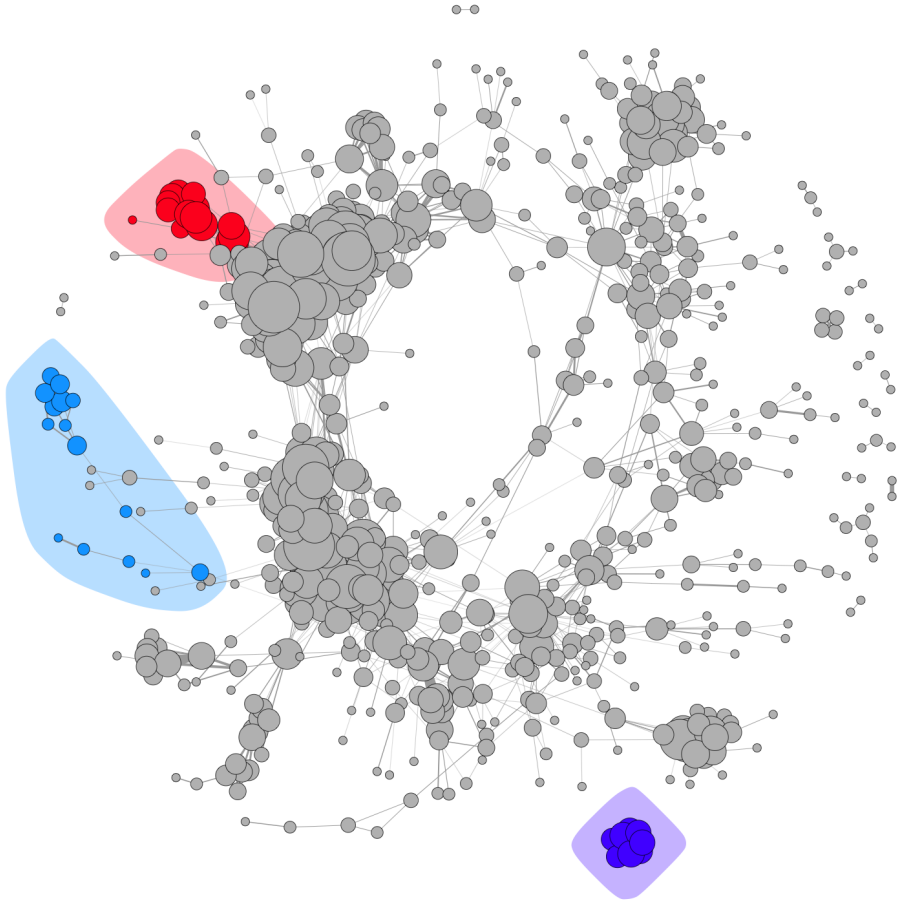
· Overlapping cancer driver genes

3 genes in 3 modules

1 gene in 8 modules

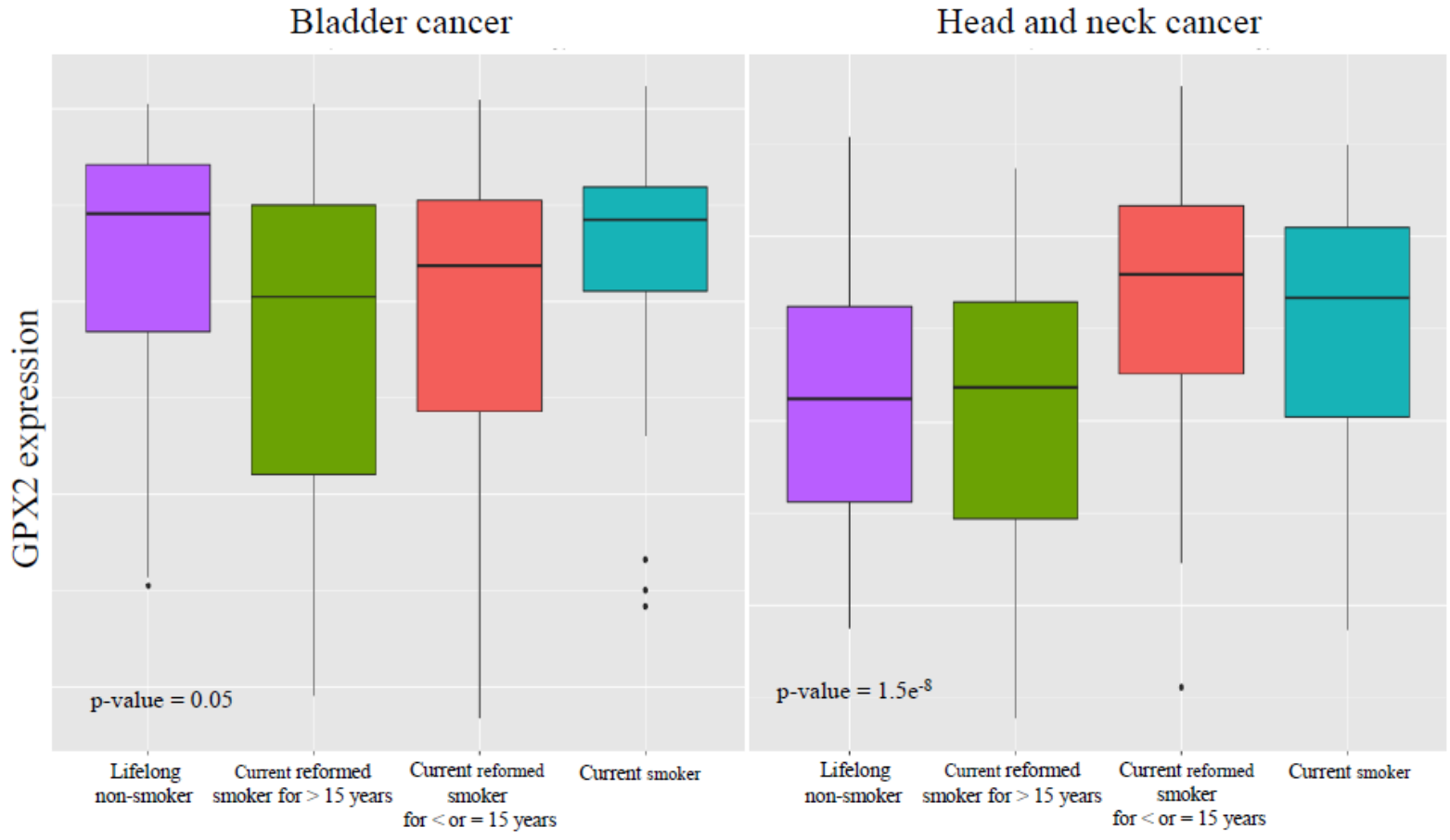
GPX2

· Enrichment in smoking related pathways

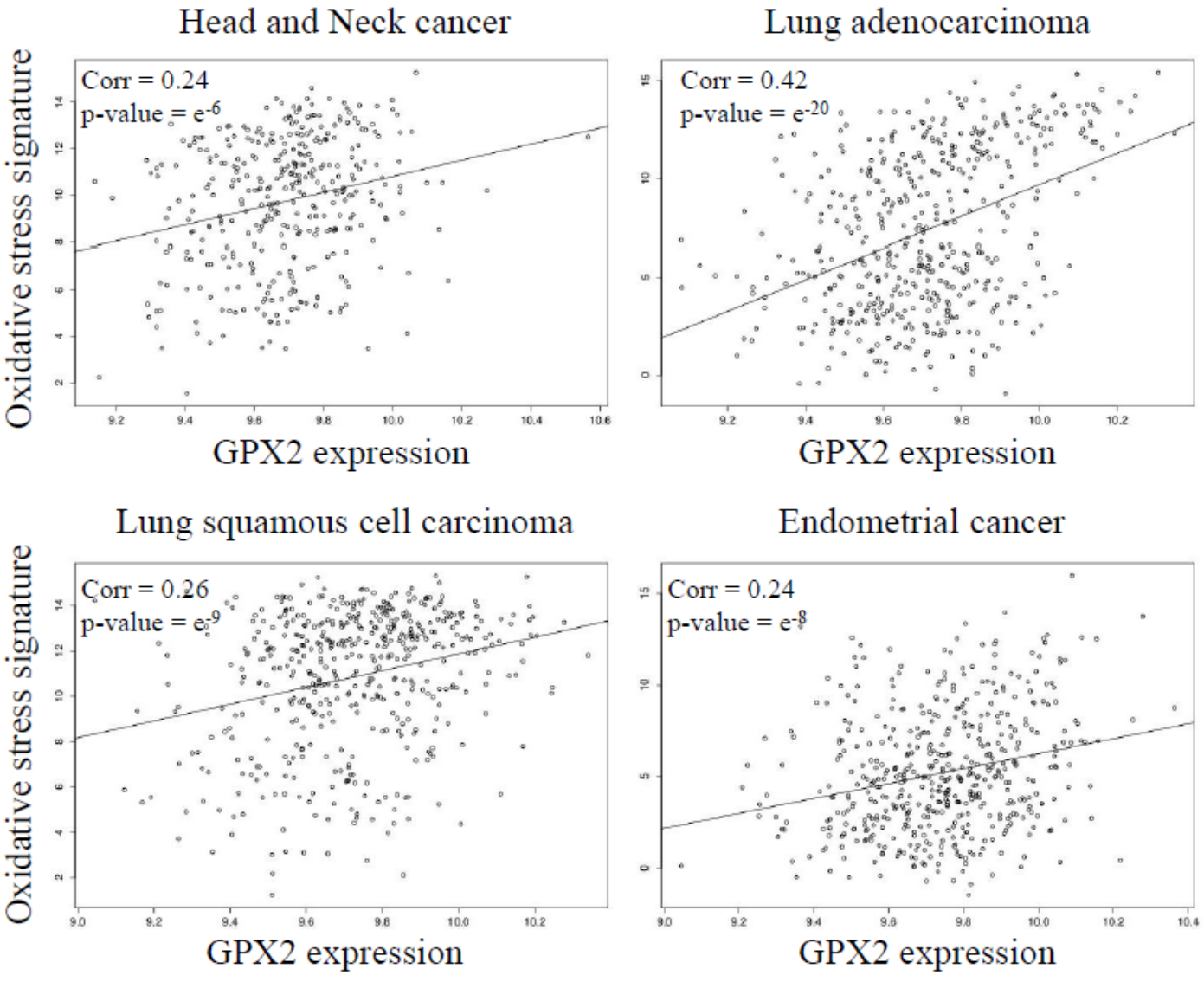


Pancancer smoking community

- Collecting clinical data, GPX2 expression is significantly associated with smoking profile.



Pancancer smoking community



Pancancer immune response community

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

· Overlapping cancer driver genes

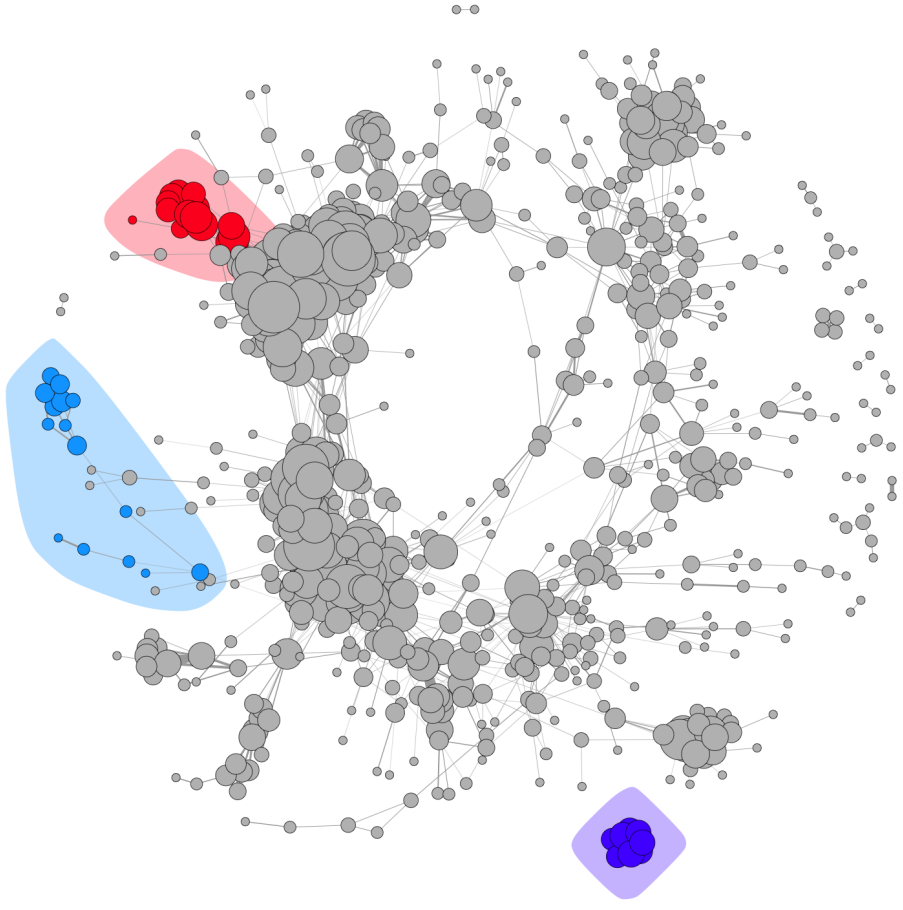
6 genes in 4 modules

1 gene in 6 modules

1 gene in 10 modules

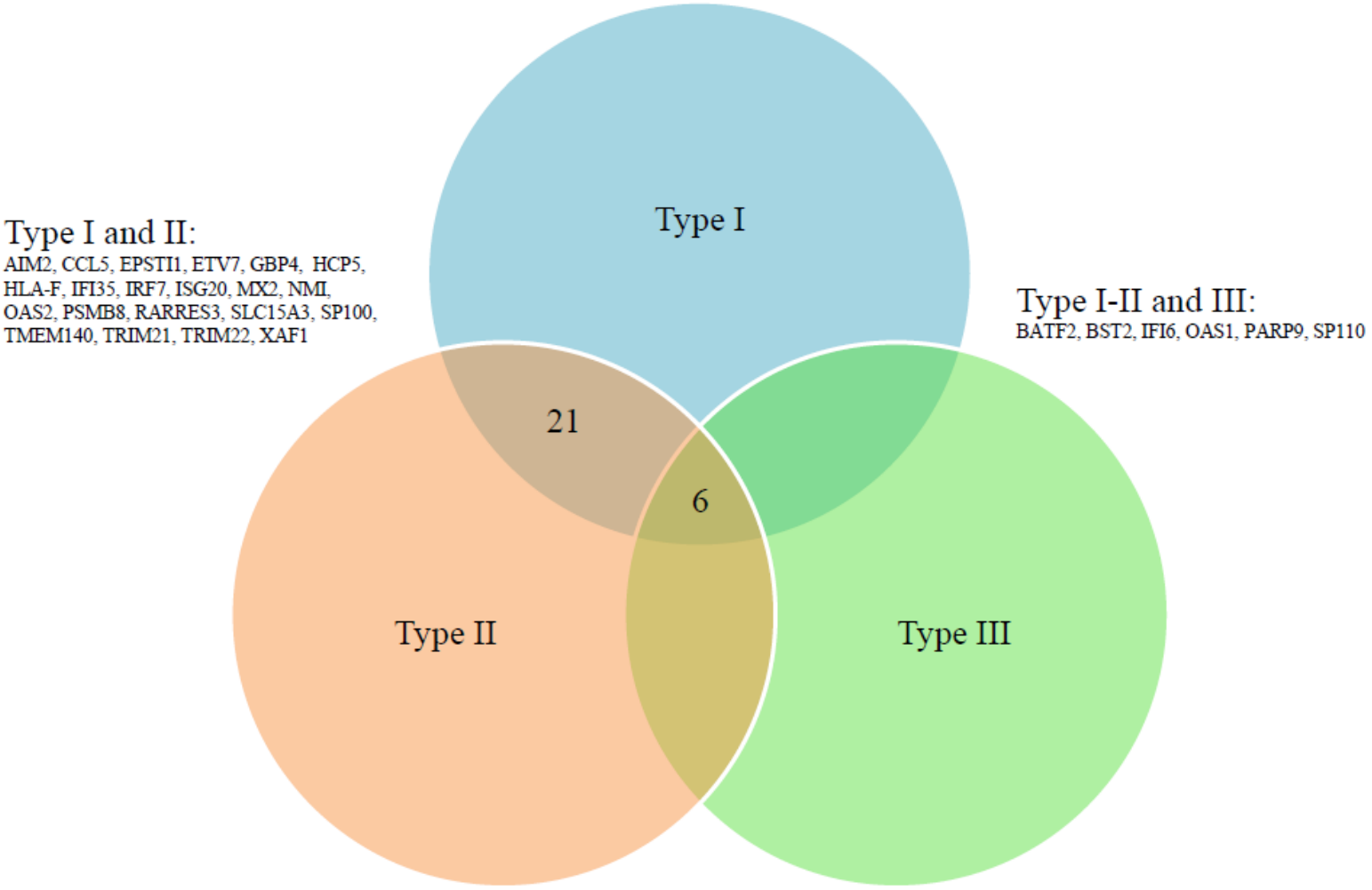
OAS2

· Enrichment in immune response pathways



Pancancer immune response community

- Most of the drivers are part of interferons.



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 mulitple 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 !

- **M. Champion**, K. Brennan, A. Gentles, T. Croonenborghs, N. Pochet, O. Gevaert. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. Soumis.
- O. Geveart, V. Villalobos, B.I. Sikic & S.K. Plevritis. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface Focus 3 :20130013, 2013.
- O. Gevaert, R. Tibshirani & S.K. Plevritis. Pancancer analysis of DNA methylation-driven genes using MethylMix. Genome Biology 16 :17, 2015.
- 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.
- D. Bell et al. Integrated genomic analyses of ovarian carcinoma. Nature 474 :609-61510.1038/nature10166, 2011.
- M.E. Newman & M. Girvan. Finding and evaluating community structure in networks. Physical review E. 69, 026113, 2004.