About OMICS International

OMICS International through its Open Access Initiative is committed to make genuine and reliable contributions to the scientific community. OMICS International hosts over 700 leading-edge peer-reviewed **Open Access Journals** and organizes over 1000+ **International Conferences** annually all over the world. OMICS International journals have over 3 million readers and the fame and success of the same can be attributed to the strong editorial board which contains over 50000 About OMICS International eminent personalities that ensure a rapid, quality and quick review process. OMICS International signed an agreement with more than 1000 **International Societies** to make healthcare information Open Access. OMICS International Conferences make the perfect platform for global networking as it brings together renowned speakers and scientists across the globe to a most exciting and memorable scientific event filled with much enlightening interactive sessions, world class exhibitions and poster presentations.



6th International Conference on Biomarkers and Clinical Research
Toronto, Canada

A comprehensive pipeline for biomarker discovery using circulating miRNA microarray data

Elena Landoni

University of Milan
Department of Clinical Sciences and Community Health

Fondazione IRCCS Istituto Nazionale dei Tumori

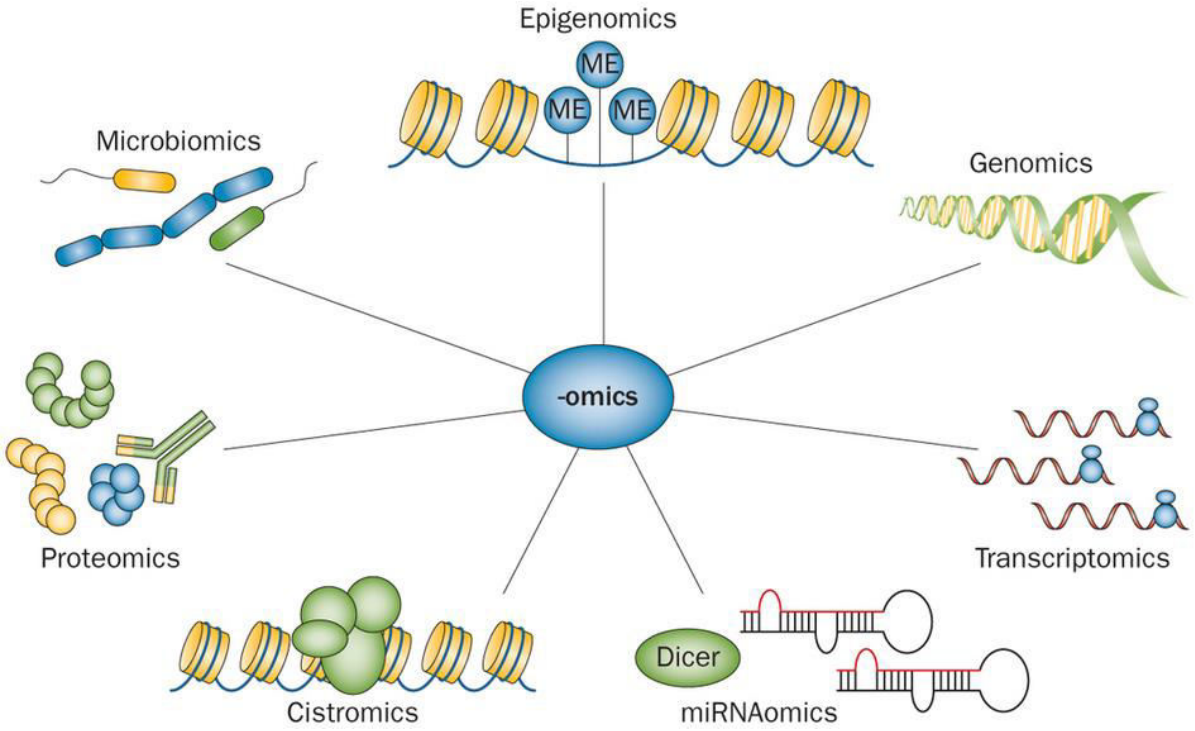
September 2nd, 2015

Outline

- Introduction: high-dimensional ‘omics’ data
- Methodological proposal
- A case study: the 4-HPR INT study
- Discussion and future perspectives



High-dimensional 'omics' data



EDERA PROJECT

Early DEtection and Risk Assessment

'Tumor microenvironment-related changes
as new tools for early detection and assessment
of high-risk diseases

http://www.ederaproject.it

FONDAZIONE IRCCS
ISTITUTO NAZIONALE
DEI TUMORI

AI4RC

Introduction

- **microRNA (miRNAs)** = highly conserved single-stranded small RNA molecules with **key role** in post-transcriptional **gene regulation** and functional implication in **cancer development and progression**

- To date more than **2600 human miRNAs** have been **identified**



- Independent studies have reported the feasibility of using plasma **miRNAs as molecular tools** for detection, prognosis and treatment decision making of various cancers

HOWEVER



Introduction (2)

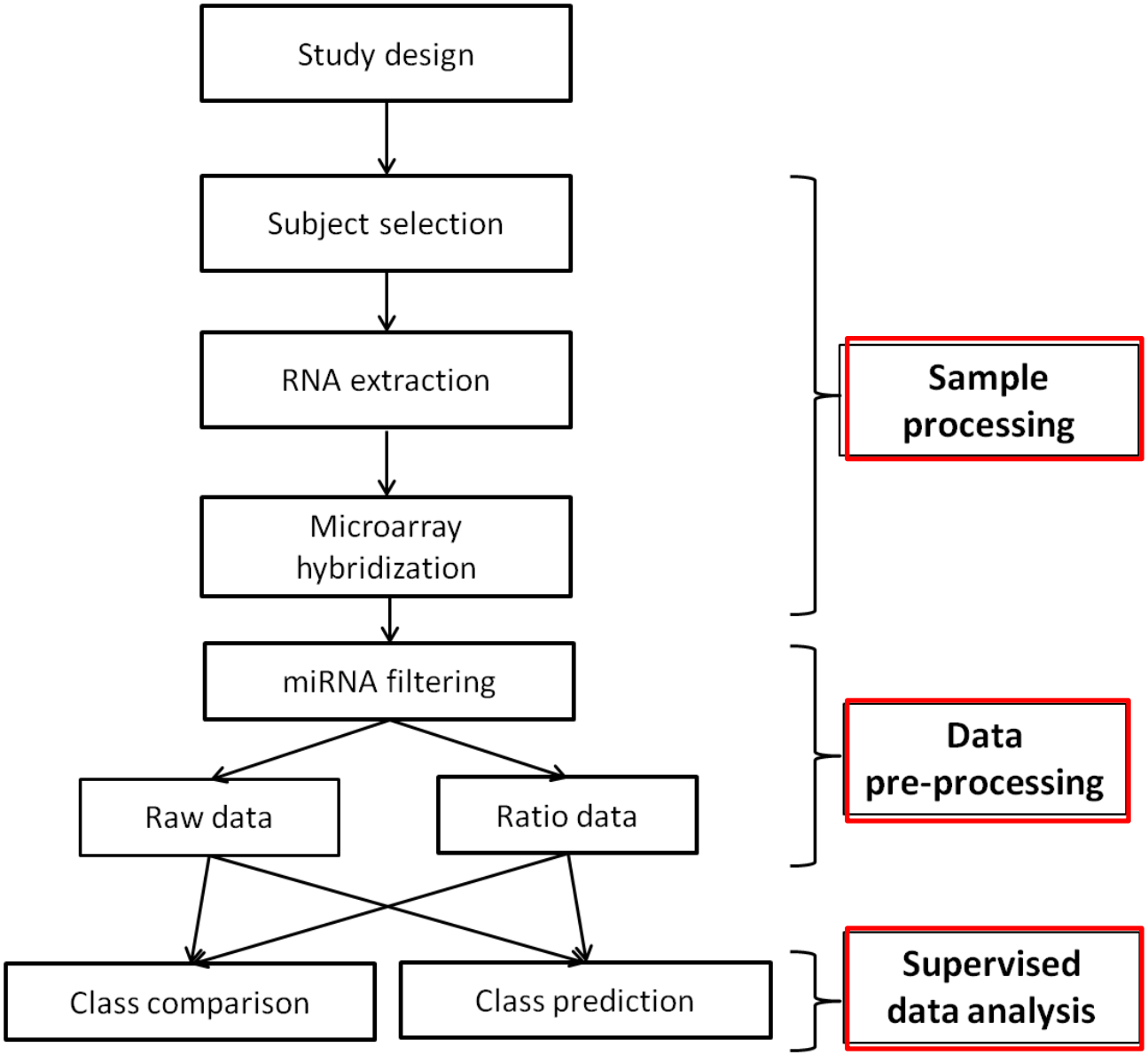
- Some **biological and technical challenges** could limit the development of miRNAs as biomarkers

AIM

Develop a multimarker hemolysis classifier using plasma miRNA data



Workflow of the strategy



Study design



- Subset of patient samples from a **randomized breast cancer prevention Trial** (1476 patients – control arm)
- Extracted RNA subjected to Agilent miRNA hybridization array -> choice of the **Agilent system** as microarray platform (**1205 human miRNAs**)
- Hemolysis evaluated with the **Hemolysis Score (HS)*** -> miRNA expression profiles from **hemolyzed** (HS > 0.14) **vs not-hemolyzed** (HS ≤ 0.057) plasma samples
- **1:2 nearest neighbour matching** (PS calipers) using disease status, age at drawing and drawing year -> **training and validation set**

* Appierto, V., Callari, M., Cavadini, E., Morelli, D., Daidone, M., Tiberio, P.: A lipemia-independent nanodrop(®)-based score to identify hemolysis in plasma and serum samples. Bioanalysis 6, 1215-1226 (2014)

Methods: data pre-processing



- Each miRNA targeted by one to four different probes and each probe spotted 10 to 40 times on the array
- The total signal for each miRNA obtained by summing the probe signals derived from Agilent Feature Extraction software
- Data **$\log_2$ transformed**, **filtered at 90%** and **normalized** with the **ratio-based approach***
- **Raw and ratio-normalized data**

Boeri, M., Verri, C., Conte, D., Roz, L., Modena, P., Facchinetti, F., et al.: MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci U S A 108, 3713-3718 (2011)

Methods: class comparison

- **Two-sample location tests** are commonly used as **univariable feature ranking** methods

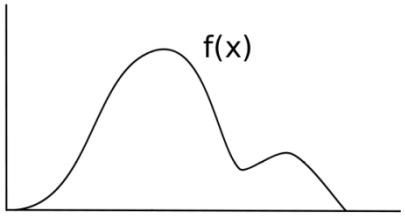
- **Non normal** (skewed, heavy tailed, multimodal) **distributions** (after data pre-processing) **+ small sample sizes** -> **location tests not adequate** to identify DE features

- **More general distributional discrepancies** might hide differential biological processes



DE = Differentially Expressed

Methods: class comparison (2)



- **t-test** -> **only location differences**
- **Anderson-Darling (AD) test** -> **'omnibus test'**, *i.e.* it considers the whole feature distribution, granting more importance to the observations in the tails
- **Benjamini-Hochberg (FDR) method** used to adjust t- and AD p-values
- **Combined use of** the **t- and** the nonparametric **AD test -> significantly DE** = FDR-adjusted p-value below the 5% level for at least one of the two tests

FDR = False Discovery Rate

Methods: class prediction

STEP 1: BOOTSTRAP SELECTION

- **B bootstrap samples** drawn from the original dataset (**B = 1000**)

- **Feature ranking** -> **JOINT** selection by **3 classification algorithms**

Algorithm	Category	High dimensionality	Complex patterns	Computational speed	Correlated features
PAM	centroid-based	✓		✓	
RF (Boruta)	tree-based	✓	✓	✓	
Elastic SCAD SVM	SVM-based	✓	✓		✓

- Graphical summarization of results by an **egg-shaped plot** (bootstrap occurrences & co-occurrences)

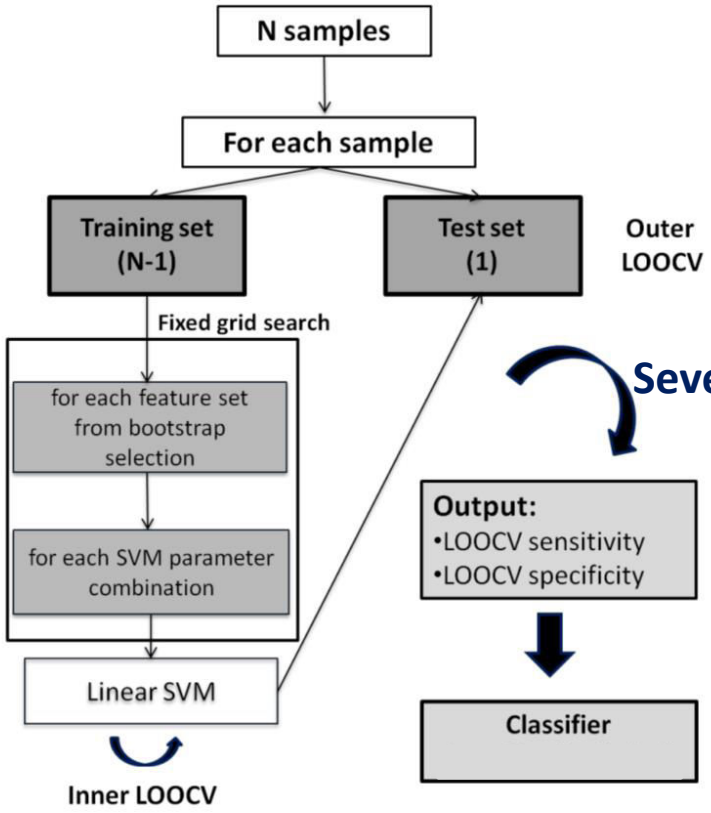
PAM = Prediction Analysis for Microarray
RF = Random Forest (Boruta feature selection algorithm)
SCAD = Elastic smoothly clipped absolute deviation

Methods: class prediction (2)

STEP 2: CLASSIFIER DEVELOPMENT



SVM parameters and number of features optimization



Final model chosen according to

1. **highest Youden index**
2. **smallest number of features**

❑ Graphical summarization of results by a **'ROC space' plot** (classification performance for all SVM models)

Results: data pre-processing

78 samples (26 hemolyzed + 52 not-hemolyzed)



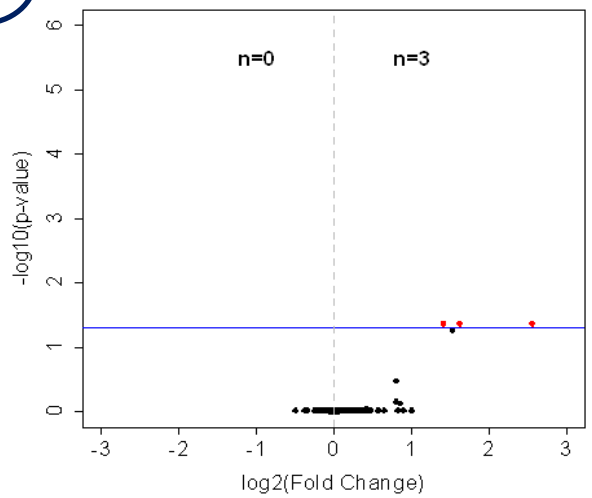
TRAINING SET: 39 samples (13 hemolyzed + 26 not-hemolyzed)

VALIDATION SET: 39 samples (13 hemolyzed + 26 not-hemolyzed)

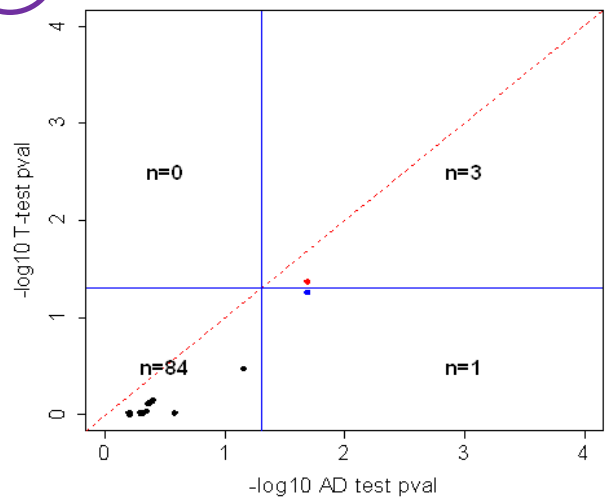
88 miRNAs (3828 ratios)

Results: class comparison

A



B

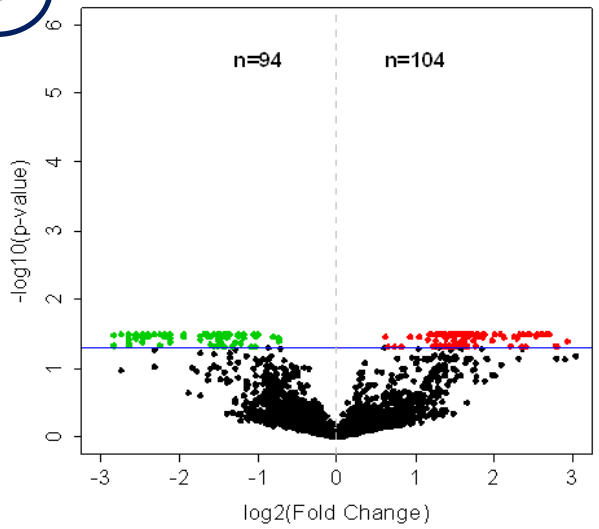


RAW DATA

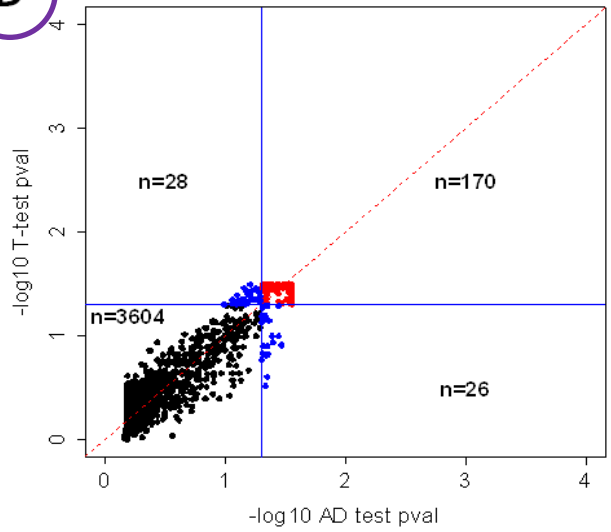
t-test

t-test
+
AD test

C



D



RATIO DATA

Results: class prediction

STEP 1: BOOTSTRAP SELECTION

BOOTSTRAP OCCURRENCES

rank	miRNA	N
1	miR-451	846
2	miR-16	779
3	miR-486-5p	734
4	miR-92a	668
5	miR-22	448
6	miR-4306	209
7	miR-21	208
8	miR-575	208
9	miR-4284	187
10	miR-202	181
11	miR-33b*	172
12	miR-223	160
13	miR-24	142
14	miR-3648	132
15	miR-630	117
16	miR-663	98
17	miR-4314	91
18	miR-187*	78
19	miR-1305	70
20	miR-3196	65
21	miR-320b	62
22	miR-3663-3p	55
23	miR-134	54
24	miR-642b	54
25	miR-548q	53
26	miR-551b*	53
27	miR-188-5p	46
28	miR-4271	46
29	miR-4257	45
30	miR-1203	42
31	miR-3125	42
32	miR-631	41
33	miR-671-5p	41
34	miR-1275	40
35	miR-3138	40

BOOTSTRAP CO-OCCURRENCES

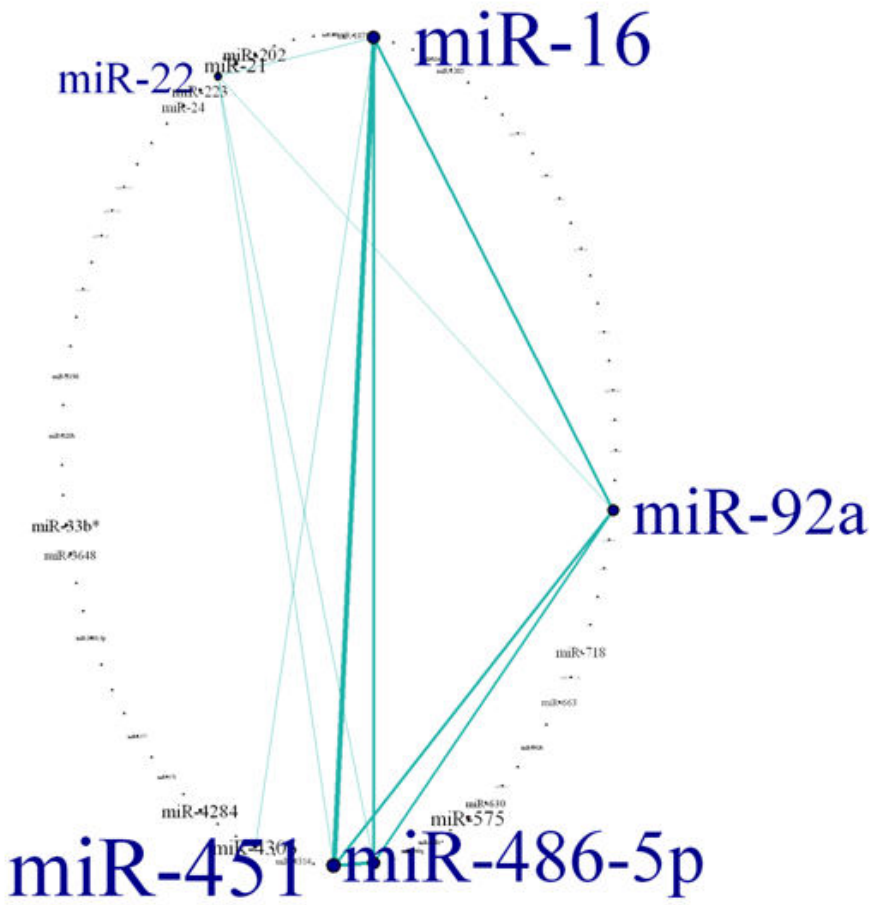
miRNA 1	miRNA 2	N
miR-16	miR-451	711
miR-451	miR-486-5p	624
miR-451	miR-92a	606
miR-16	miR-92a	604
miR-16	miR-486-5p	587
miR-486-5p	miR-92a	554
miR-16	miR-22	411
miR-451	miR-22	402
miR-92a	miR-22	397
miR-485-5p	miR-22	344

Results: class prediction

STEP 1: BOOTSTRAP SELECTION REPRESENTATION

EGG-SHAPED PLOT

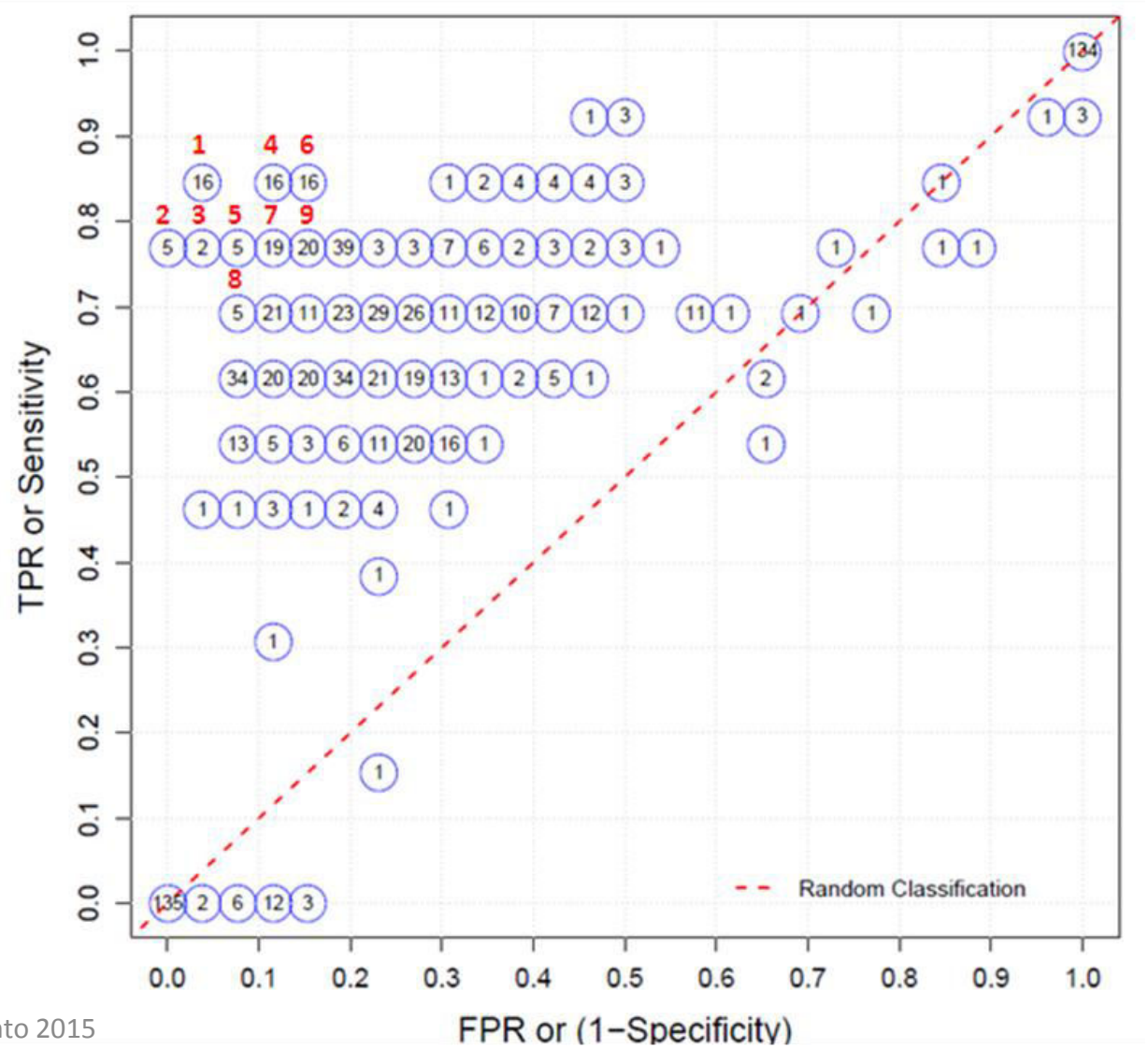
BOOTSTRAP OCCURRENCES & CO-OCCURRENCES



Results: class prediction (2)

STEP 2: CLASSIFIER DEVELOPMENT REPRESENTATION

ROC SPACE PLOT



RAW DATA

Results: class prediction (3)

Classification performance of the best performing groups of models evaluated in the training set					Parameters of the chosen model			Classification performance of the chosen model evaluated in the validation set		
Group ID*	N models	Sens	Spec	Youden	N miR	SVM Cost	SVM Weights	Sens [CI]	Spec [CI]	Youden [CI]
						Raw data				
1	16	0.85	0.96	0.81	35	10	(0.5; 0.5)	0.77 [0.54-0.92]	0.77 [0.61-0.92]	0.54 [0.23-0.81]
2	5	0.77	1.00	0.77	35	1	(0.5; 0.5)	0.85 [0.61-1.00]	0.81 [0.65-0.92]	0.65 [0.38-0.85]
3	2	0.77	0.96	0.73	30	1	(0.5; 0.5)	0.85 [0.61-1.00]	0.85 [0.69-0.96]	0.69 [0.42-0.88]
4	16	0.85	0.88	0.73	40	10	(0.5; 0.5)	0.77 [0.54-0.92]	0.73 [0.54-0.88]	0.50 [0.19-0.77]
5	5	0.77	0.92	0.69	35	1	(0.4; 0.6)	0.85 [0.61-1.00]	0.73 [0.54-0.88]	0.58 [0.31-0.81]
6	16	0.85	0.85	0.69	50	10	(0.5; 0.5)	0.85 [0.61-1.00]	0.69 [0.50-0.88]	0.54 [0.27-0.81]
7	19	0.77	0.88	0.65	40	1	(0.4; 0.6)	0.77 [0.54-0.92]	0.69 [0.50-0.85]	0.46 [0.15-0.73]
8	5	0.69	0.92	0.61	3	100	(0.4; 0.6)	0.77 [0.54-0.92]	0.96 [0.88-1.00]	0.73 [0.46-0.92]
9	20	0.77	0.85	0.61	5	10	(0.4; 0.6)	0.69 [0.46-0.92]	0.92 [0.81-1.00]	0.61 [0.35-0.85]
						Ratio data				
1	1	0.92	0.81	0.73	500 (88)	0.01	(0.2; 0.8)	0.92 [0.77-1.00]	0.65 [0.46-0.85]	0.58 [0.31-0.81]
2	1	0.77	0.92	0.69	17 (16)	0.01	(0.3; 0.7)	0.77 [0.54-0.92]	0.92 [0.81-1.00]	0.69 [0.42-0.92]
3	1	0.69	1.00	0.69	90 (50)	0.01	(0.5; 0.5)	0.69 [0.38-0.92]	1.00 [1.00-1.00]	0.69 [0.38-0.92]
4	2	0.85	0.85	0.69	150 (66)	0.01	(0.2; 0.8)	0.92 [0.77-1.00]	0.69 [0.50-0.85]	0.61 [0.38-0.81]
5	35	0.69	0.96	0.65	4 (5)	0.1	(0.5; 0.5)	0.77 [0.54-0.92]	1.00 [1.00-1.00]	0.77 [0.54-0.92]
6	4	0.77	0.88	0.65	500 (88)	0.01	(0.4; 0.6)	0.92 [0.77-1.00]	0.77 [0.58-0.92]	0.69 [0.46-0.88]
7	3	0.85	0.81	0.65	600 (88)	0.01	(0.2; 0.8)	0.92 [0.77-1.00]	0.65 [0.46-0.85]	0.58 [0.31-0.81]
8	11	0.61	1.00	0.61	2 (3)	0.1	(0.5; 0.5)	0.77 [0.54-0.92]	1.00 [1.00-1.00]	0.77 [0.54-0.92]
9	23	0.69	0.92	0.61	3 (4)	0.1	(0.4; 0.6)	0.77 [0.54-0.92]	0.96 [0.88-1.00]	0.73 [0.50-0.92]
10	54	0.77	0.85	0.61	4 (5)	0.1	(0.3; 0.7)	0.77 [0.54-0.92]	0.88 [0.73-1.00]	0.65 [0.38-0.88]
11	18	0.61	0.96	0.58	3 (4)	0.1	(0.5; 0.5)	0.77 [0.54-0.92]	1.00 [1.00-1.00]	0.77 [0.54-0.92]
12	59	0.69	0.88	0.58	2 (3)	10	(0.4; 0.6)	0.77 [0.54-0.92]	1.00 [1.00-1.00]	0.77 [0.54-0.92]

Leave-One-Out Cross Validation

Discussion

- We developed a **general analysis strategy** to deal with some **issues** arising in the **supervised analyses** of **plasma miRNA** from **hybridisation array data**

DATA PRE-PROCESSING

- ✓ **Ratio method** chosen **as normalization method -> automatic handling** of a specific normalizer for each DE miRNA

Discussion (2)

CLASS COMPARISON

- ✓ **Location tests** such as the **t-test** are commonly applied **(univariable statistical analyses) BUT normality** is often **not fulfilled** with **plasma miRNA data** (**skewed, heavy-tailed or multimodal distributions**), especially if associated with **small sample sizes** -> t- + AD test -> discover of miRNAs discarded by the t-test

- ✓ **The AD test** is particularly **valuable** when distributions differ in the **tails**, which could reveal **underlying biological differences**

- ✓ **Caution** in using **class comparison** for **ranking purposes**

Discussion (3)

CLASS PREDICTION

- ✓ **Bootstrap selection** with the application of **3 machine learning algorithms** (**PAM, RF, Elastic SCAD SVM**) + **classifier cross validation** -> **robust** results

- ✓ **Egg-shaped plot** used as a tool for giving an **insight** of **interconnections among the selected features** (inherent biological dependences)

- ✓ **Different SVM models <->** different subgroups of features <-> **choice** of **parsimonious models**

- ✓ Caution due to **small sample size (large bootstrap intervals)**

Discussion (5)

- ✓ **5 top miRNAs: miR-486-5p, miR-92, miR-451, miR-16** (literature) **+ miR-22** (new one)

- ✓ **Consistent results** in **class comparison** and **bootstrap selection** (raw and ratio data)

- ✓ The **concordance of the results with literature data** corroborated the ability of **HS** as **pre-analytic hemolysis detector**

- ✓ **Study limitations**:
  1. **lack of availability** of a **testing set** for unbiased assessment of classifier performance
  2. **SVM** -> **'surrogated'** class probabilities

- ✓ Application to **other types** of **'omics' data** (with proper adjustments)

Future perspectives

❑ DATA PRE-PROCESSING

- Search for an **alternative normalization method**

❑ CLASS COMPARISON

- Use of the **AD test** and search for a corresponding effect measure **-> volcano plots** able to detect as **DE also features** with **irregular** expression **distributions**

❑ CLASS PREDICTION

- Fitting of **other types** of **models** for **classifier development**

Acknowledgements

Gabriella Sozzi — *PI EDERA PROJECT*

Luigi Mariani
Rosalba Miceli — *Scientific direction*

Medical Statistics, Biometry and Bioinformatics Unit
Clinical Epidemiology and Trial Organization Unit

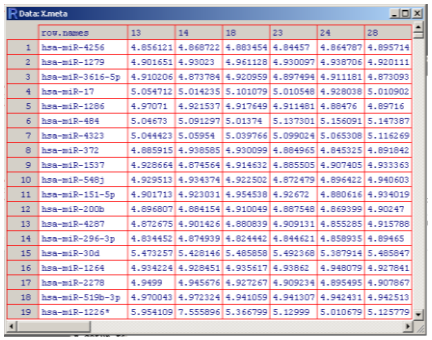
Valentina Angeloni
Valentina Appierto
Maurizio Callari
Maria Grazia Daidone
Paola Tiberio — *Department of Experimental Oncology and Molecular Medicine*

Biomarkers Unit

FONDAZIONE IRCCS ISTITUTO NAZIONALE DEI TUMORI

http://www.ederaproject.it

High-dimensional 'omics' data: issues



	row.names	13	14	18	23	24	28
1	hsa-miR-4256	4.856121	4.868722	4.883454	4.84457	4.864787	4.895714
2	hsa-miR-1279	4.901651	4.93023	4.961128	4.930097	4.938706	4.920111
3	hsa-miR-3616-5p	4.910206	4.873784	4.920959	4.897494	4.911181	4.873093
4	hsa-miR-17	5.054712	5.014235	5.101079	5.010548	4.928038	5.010902
5	hsa-miR-1286	4.97071	4.921537	4.917649	4.911481	4.88476	4.89716
6	hsa-miR-484	5.04673	5.091297	5.01374	5.137301	5.156091	5.147387
7	hsa-miR-4323	5.044423	5.05954	5.039766	5.099024	5.065308	5.116269
8	hsa-miR-372	4.885915	4.938585	4.930099	4.884965	4.845325	4.891842
9	hsa-miR-1537	4.928664	4.874564	4.914632	4.885505	4.907405	4.933363
10	hsa-miR-548j	4.929513	4.934374	4.922502	4.872479	4.896422	4.940603
11	hsa-miR-151-5p	4.901713	4.923031	4.954538	4.92672	4.880616	4.934019
12	hsa-miR-200b	4.896807	4.884154	4.910049	4.887548	4.869399	4.90247
13	hsa-miR-4287	4.872675	4.901426	4.880839	4.909131	4.855285	4.915788
14	hsa-miR-296-3p	4.834452	4.874939	4.824442	4.844621	4.858935	4.89465
15	hsa-miR-30d	5.473257	5.428146	5.485858	5.492368	5.387914	5.485847
16	hsa-miR-1264	4.934224	4.928451	4.935617	4.93862	4.948079	4.927841
17	hsa-miR-2278	4.9499	4.945676	4.927267	4.909234	4.895495	4.907867
18	hsa-miR-519b-3p	4.970043	4.972324	4.941059	4.941307	4.942431	4.942513
19	hsa-miR-1226*	5.954109	7.555896	5.366799	5.12999	5.010679	5.125779

***n** samples*
<<
***p** features*

DATA PREPARATION	CLASS COMPARISON	CLASS PREDICTION
choice of the proper data normalization method	search for differentially expressed (DE) features	screening of features for classifier development

The 4-HPR INT study

ABSTRACT

Circulating miRNAs have the potential as cancer biomarkers but **no consolidated guidelines** are established **for discovery analyses**. Several issues (*e.g.* data normalization, expected miRNA up regulation in one of classes, sample size limitation) can affect results making many approaches unsuitable. We developed a **structured pipeline** with innovative applications of existing bioinformatics methods including: **1) an assumption-independent normalization method** based on **miRNA ratios** in data pre processing; **2) the combination** of the results **of two statistical tests (t- and Anderson Darling)** to detect miRNAs with significant fold change or general distributional differences **in class comparison**; **3)** the application of **a bootstrap selection procedure together with machine learning techniques** to guarantee result generalizability and study the inter connections among the selected miRNAs **in class prediction**. We applied the pipeline to **compare hemolized and non hemolized plasma samples**, identifying **four miRNAs** known to be **hemolysis related (miR 486-5p, miR-92a, miR-451, miR-16) together with a new one, miR-22.**

Team Biomarkers welcomes you all to the next chapter – 7th International Conference on Biomarkers & Clinical Research scheduled for Nov 28-30, 2016 in Baltimore, USA

Please Visit:

www.conferenceseries.com

http://www.omicsonline.org/

http://biomarkers.conferenceseries.com/

Let Us Meet Again in Baltimore, USA