A three-state model for multidimensional data integration

Claudia Rangel-Escareño

Head Computational Genomics

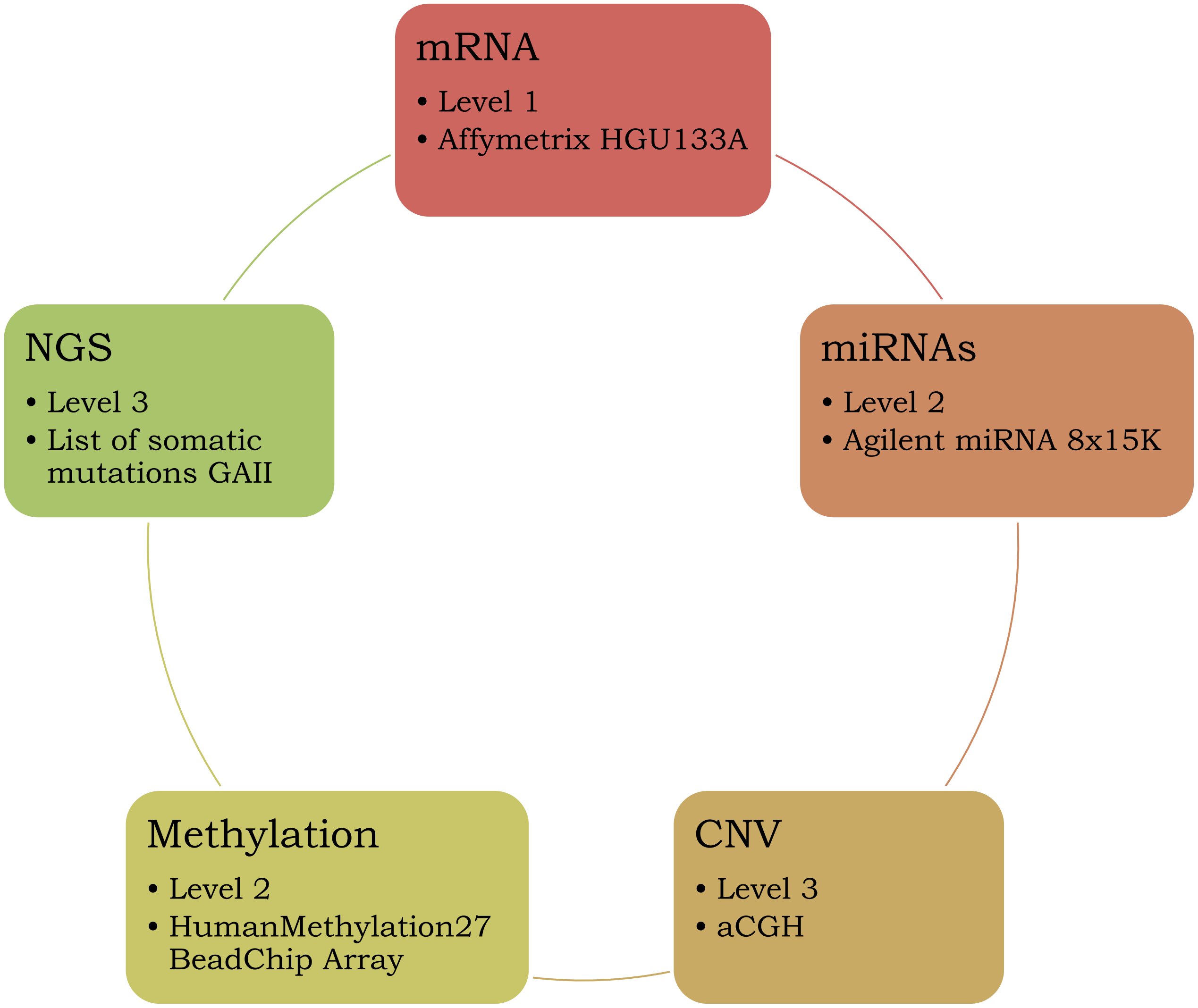
INMEGEN

National Institute of Genomic Medicine
MEXICO

Outline

- GBM Data
- Low-level analyses
- Pipeline and timeline
- Access to TCGA data
- Data integration – real challenge
- Strategy for integrative analysis
- Results
- Discussion
- Remarks

GBM Data



mRNA
- Level 1
- Affymetrix HGU133A

miRNAs
- Level 2
- Agilent miRNA 8x15K

CNV
- Level 3
- aCGH

Methylation
- Level 2
- HumanMethylation27 BeadChip Array

NGS
- Level 3
- List of somatic mutations GAII

Low-level Analysis

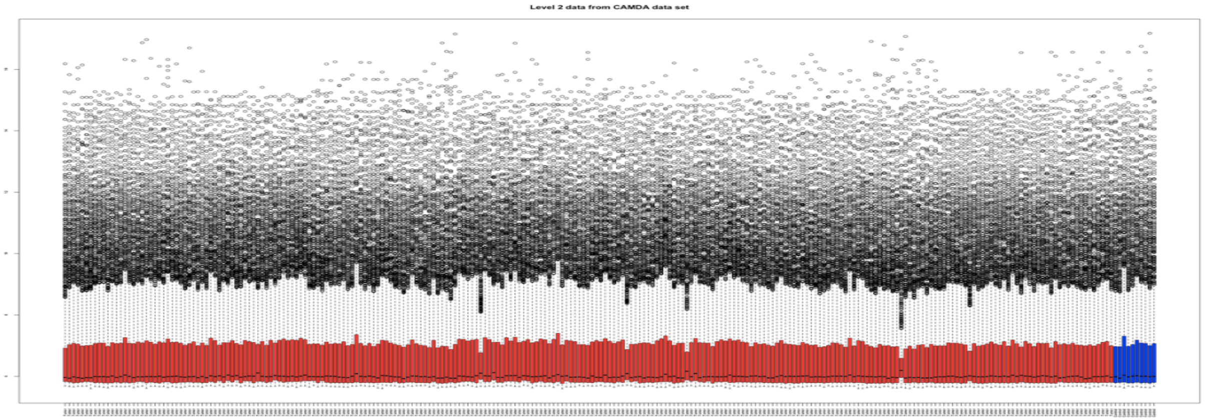
mRNA level 1 data

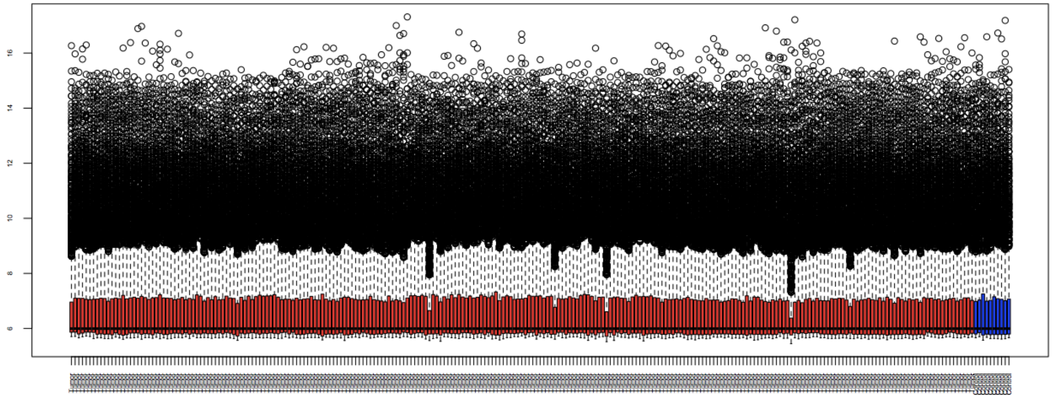
- 495 tumor samples and 10 controls
- normalized using quantile normalization
- summarized using medianpolish
- classification based on log fold-change, B-statistic and adjusted p-values

Low-level Analysis

miRNA level 2 data

- 245 tumor sam- ples and 10 controls
- According to TCGA portal data were background corrected using RMA and quantile normalized.





Low-level Analysis

Methylation level 2 data

- from 291 tumor samples and 1 control with 6 replicates
- normalized and processed using genome wide Infinium HumanMethylation27 BeadChip Array
- ~ 27,578 CpG sites.
- Beta-values and confidence p-values were further examined
- Missing beta-values were calculated using the signal intensity (M) and the un-methylated signal intensity (U).

Low-level Analysis

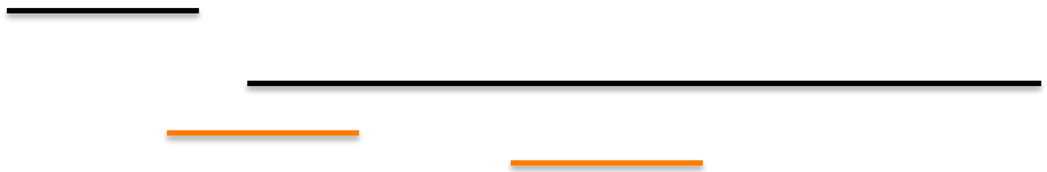
CNVs level 3 data

- Data for 461 samples processed with array CGH technology
- Data reported to be lowess normalized.
- Regions of gain and loss were identified using Circular Binary Segmentation algorithm

Our part:

- Which genes are in each reported segment?

- Algorithm



National Institute of Genomic Medicine
MEXICO

Low-level Analysis

NGS level 3 data

- somatic nucleotide alteration data for 143 samples in 3 databases were analyzed.
- The three databases were combined and relevant mutations were selected.
- The final database contained 1032 unique gene-mutation pairs, for 500 different genes and 7 different mutation types:

● Missense ● Splice_Site ● Nonsense ● Unkown

● Silent ● Frame_Shift ● In_Frame

National Institute of Genomic Medicine
MEXICO

Clinical Data

IDs curated "TA.0001.F.D.44.WT.NPG.RA.CH.B1”

- T : describes sample type (T=Primary Tumor, B=Blood Derived Normal, N=Solid Tumor Tissue)
- A : indicates replicate A=1, B=2
- "0001”: corresponds to patient ID
- F : indicates gender (F=Female, M=Male)
- D : corresponden al Vital status (D= Deceased, L=Living)
- "44” : is the patient´s age
- Cancer status: WT=with tumor,TF= tumor free
- Prior glioma: PG= Prior glioma,NPG=non-prior glioma
- Therapy: CH= chemotherapy,HO=hormonal therapy,IM= immuno therapy,RA=radiations therapy,TM= targeted molecular therapy

All Data

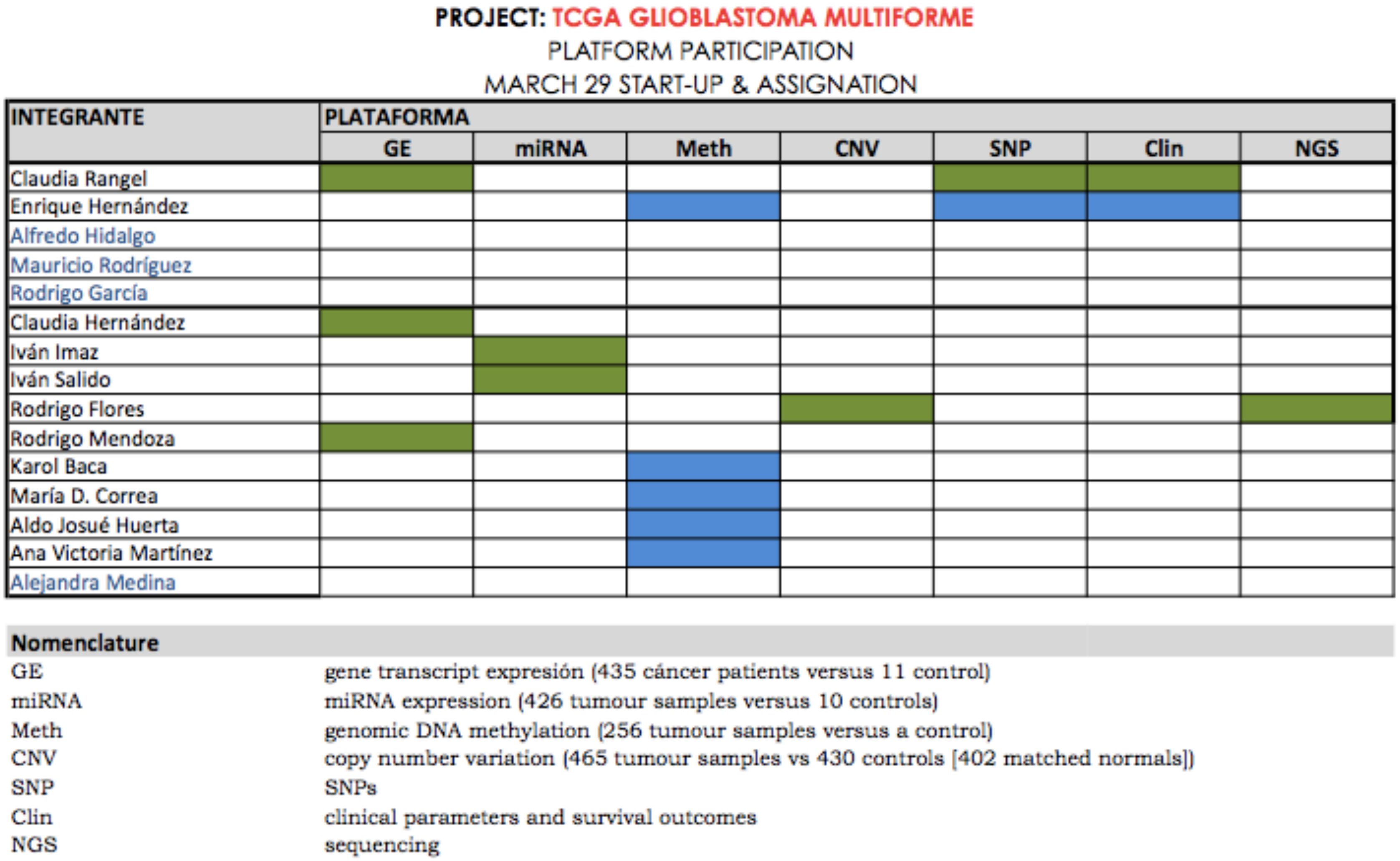
- Genome_Wide_SNP_6 --> GWS6
- HG-CGH-244A --> CGH244
- HumanHap550 --> Hh550
- HumanMethylation27 --> HMet27
- IlluminaDNAMethylation --> IllMet
- HT_HG-U133A --> Exp133A
- HuEx-1_0-st-v2 --> ExpExon
- AgilentG4502A_07 --> ExpAgi
- H-miRNA_8x15K --> ExpmiR
- ABI --> ABI
- HG-CGH-415K_G4124A --> CGH415

TCGA data portal

Notes & Remarks

- Gene expression data from three different platforms was badly combined. So can´t always trust level 3 data …
- Access to SNP 6.0 array data would have given us the opportunity of doing some ancestry analysis
- Access to Level 1 Human Gene 1.0 ST would have given us a chance to do outlier detection using COPA

How did we do it?



PROJECT: TCGA GLIOBLASTOMA MULTIFORME

PLATFORM PARTICIPATION

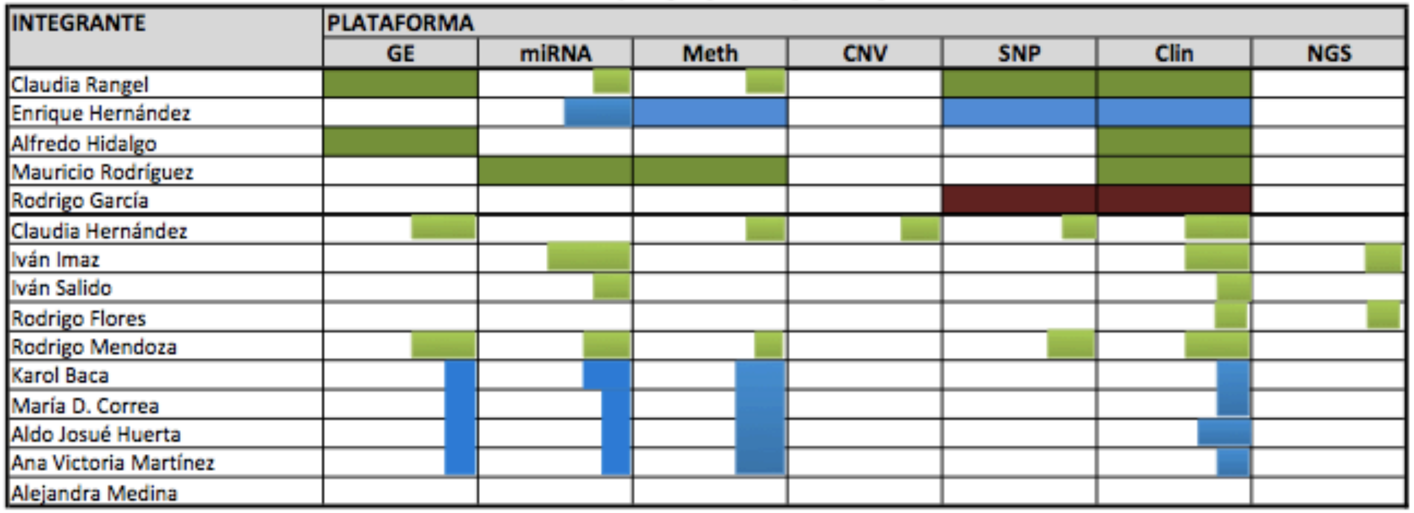
MARCH 29 START-UP & ASSIGNATION

INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩				🟩	🟩	
Enrique Hernández			🟦		🟦	🟦	
Alfredo Hidalgo							
Mauricio Rodríguez							
Rodrigo García							
Claudia Hernández	🟩						
Iván Imaz		🟩					
Iván Salido		🟩					
Rodrigo Flores				🟩			🟩
Rodrigo Mendoza	🟩						
Karol Baca			🟦				
María D. Correa			🟦				
Aldo Josué Huerta			🟦				
Ana Victoria Martínez			🟦				
Alejandra Medina							

Nomenclature	
GE	gene transcript expresión (435 cáncer patients versus 11 control)
miRNA	miRNA expression (426 tumour samples versus 10 controls)
Meth	genomic DNA methylation (256 tumour samples versus a control)
CNV	copy number variation (465 tumour samples vs 430 controls [402 matched normals])
SNP	SNPs
Clin	clinical parameters and survival outcomes
NGS	sequencing

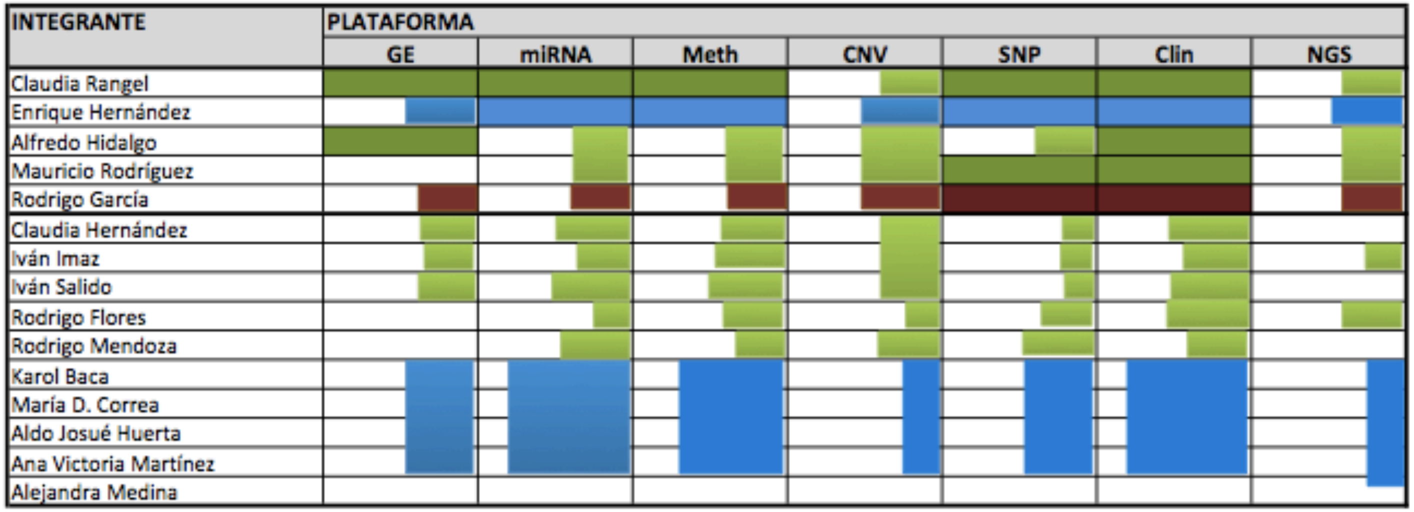
How did we do it?

APRIL 5 TO 19: RESEARCH



INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩	🟩	🟩		🟩	🟩	
Enrique Hernández		🟦	🟦		🟦	🟦	
Alfredo Hidalgo	🟩					🟩	
Mauricio Rodríguez		🟩	🟩			🟩	
Rodrigo García					🟥	🟥	
Claudia Hernández	🟩		🟩	🟩	🟩	🟩	
Iván Imaz		🟩				🟩	🟩
Iván Salido		🟩				🟩	
Rodrigo Flores						🟩	🟩
Rodrigo Mendoza	🟩	🟩	🟩		🟩	🟩	
Karol Baca	🟦	🟦	🟦			🟦	
María D. Correa	🟦	🟦	🟦			🟦	
Aldo Josué Huerta	🟦	🟦	🟦			🟦	
Ana Victoria Martínez	🟦	🟦	🟦			🟦	
Alejandra Medina							

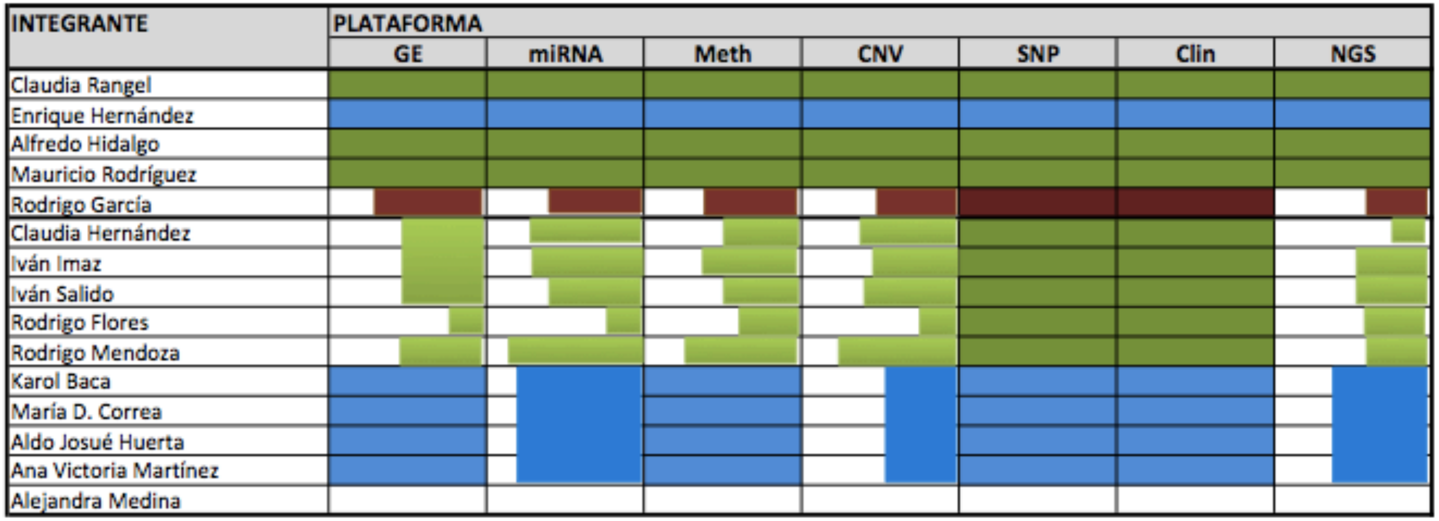
APRIL 26: RESEARCH & EXECUTION



INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Enrique Hernández	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alfredo Hidalgo	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Mauricio Rodríguez		🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo García	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Claudia Hernández	🟩	🟩	🟩	🟩	🟩	🟩	
Iván Imaz	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Salido	🟩	🟩	🟩	🟩	🟩	🟩	
Rodrigo Flores		🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Mendoza		🟩	🟩	🟩	🟩	🟩	
Karol Baca	🟦	🟦	🟦	🟦	🟦	🟦	🟦
María D. Correa	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Aldo Josué Huerta	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Ana Victoria Martínez	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alejandra Medina							

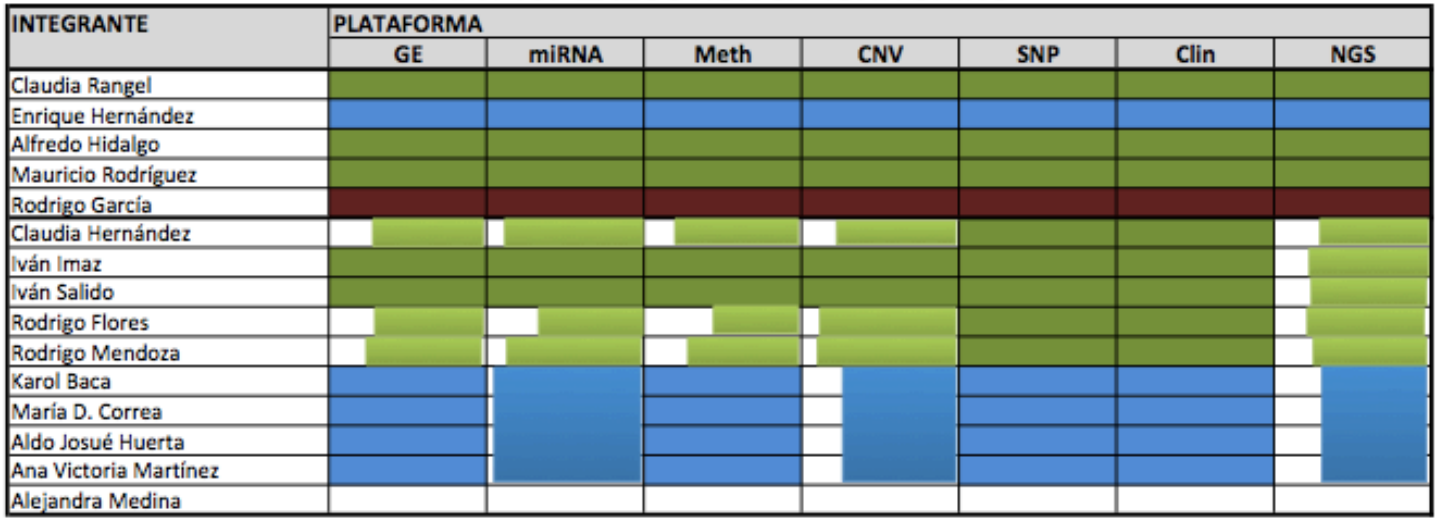
How did we do it?

MAY 3: EXECUTION & MONITORING



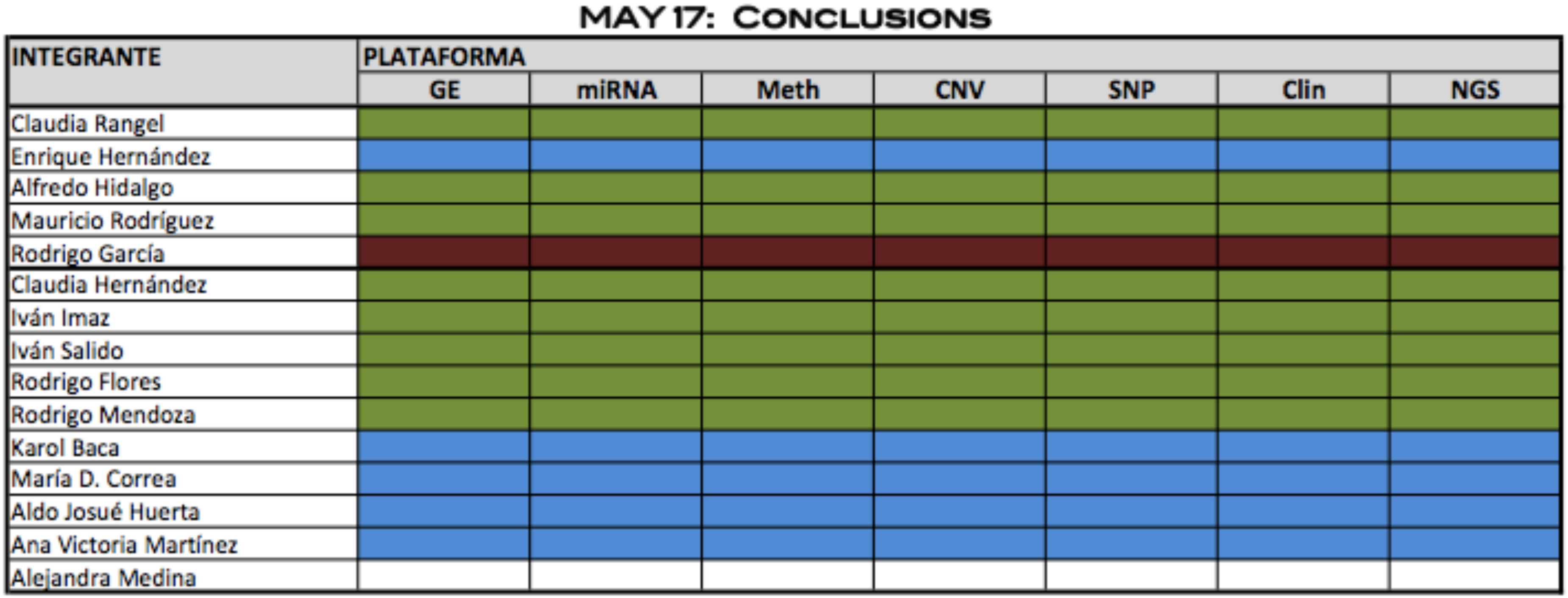
INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Enrique Hernández	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alfredo Hidalgo	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Mauricio Rodríguez	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo García	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Claudia Hernández	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Imaz	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Salido	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Flores	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Mendoza	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Karol Baca	🟦	🟦	🟦	🟦	🟦	🟦	🟦
María D. Correa	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Aldo Josué Huerta	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Ana Victoria Martínez	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alejandra Medina							

MAY 10: EXECUTION & MONITORING



INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Enrique Hernández	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alfredo Hidalgo	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Mauricio Rodríguez	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo García	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Claudia Hernández	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Imaz	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Salido	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Flores	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Mendoza	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Karol Baca	🟦	🟦	🟦	🟦	🟦	🟦	🟦
María D. Correa	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Aldo Josué Huerta	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Ana Victoria Martínez	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alejandra Medina							

How did we do it?

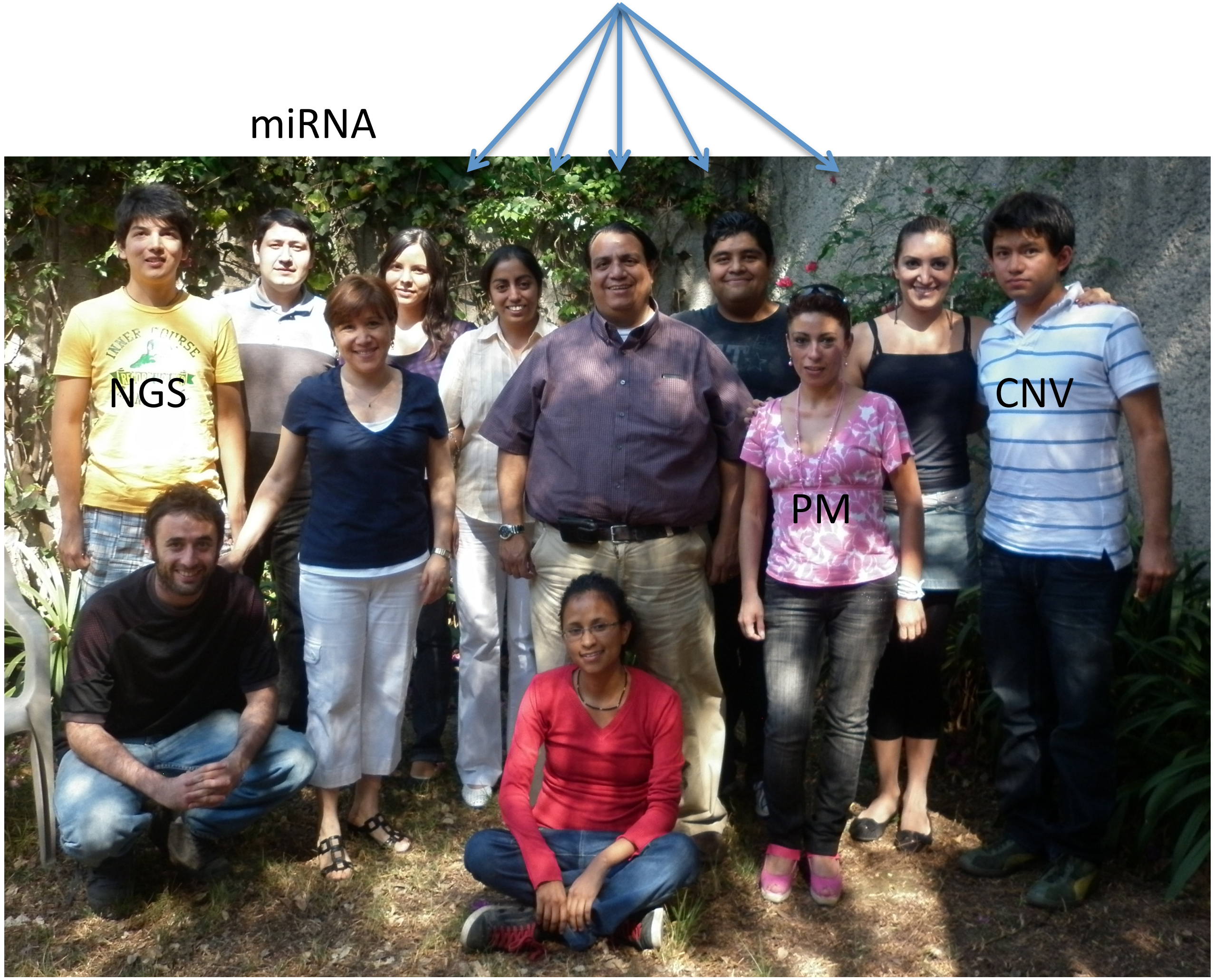


MAY 17: CONCLUSIONS

INTEGRANTE	PLATAFORMA						
	GE	miRNA	Meth	CNV	SNP	Clin	NGS
Claudia Rangel	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Enrique Hernández	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alfredo Hidalgo	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Mauricio Rodríguez	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo García	🟥	🟥	🟥	🟥	🟥	🟥	🟥
Claudia Hernández	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Imaz	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Iván Salido	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Flores	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Rodrigo Mendoza	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Karol Baca	🟦	🟦	🟦	🟦	🟦	🟦	🟦
María D. Correa	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Aldo Josué Huerta	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Ana Victoria Martínez	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Alejandra Medina							

National Institute of Genomic Medicine
MEXICO

Computational Genomics



methylation

miRNA

Strategy for integrative analysis

3-State Model

Three-State Model

- Combinatorial data driven approach

- We first selected the list for most significant genes based on mRNA levels
- For each gene *i*, let S_{i1}, S_{i2}, … , S_{ik} be a sequence of states where S_{ik} denotes the state of gene *i* in platform *k*
- Each state can take values {-1,0,1} based on whether it reports to be *up*, with *no change* or *down* regulated respectively.
- Platforms are combined following basic set theory

$$P_i \bigcup P_j = (P_i \bigcap P_j) \bigcup (P_i \setminus P_j) \bigcup (P_j \setminus P_i)$$

Three-State Model

Example

- Suppose we choose 3-Platform approach

  {Mutation, Methylation, mRNA}

- A gene taking values {1,-1,1} indicates that it contains somatic nucleotide alteration, is hypo-methylated and differentially up-regulated

How many scenarios?

- Under the approach described we have *3^k^* possible scenarios for a k-platform analysis assuming a 3-state model
- It allows simple consideration such as 2-state for NGS
- So, when we begin the integration we could have up to

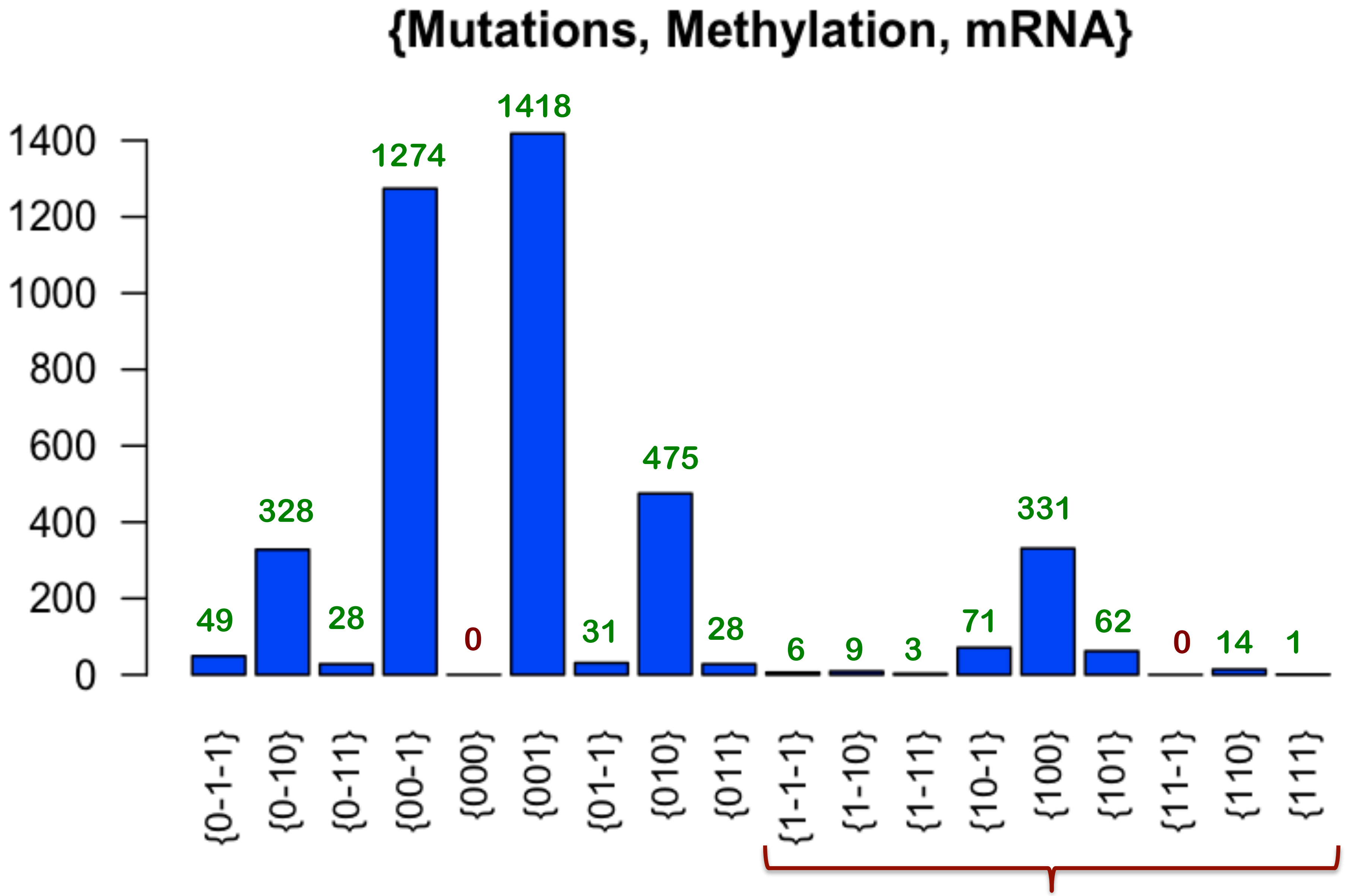
$$\sum_{h=1}^{k} 3^{h} \binom{k}{h}$$

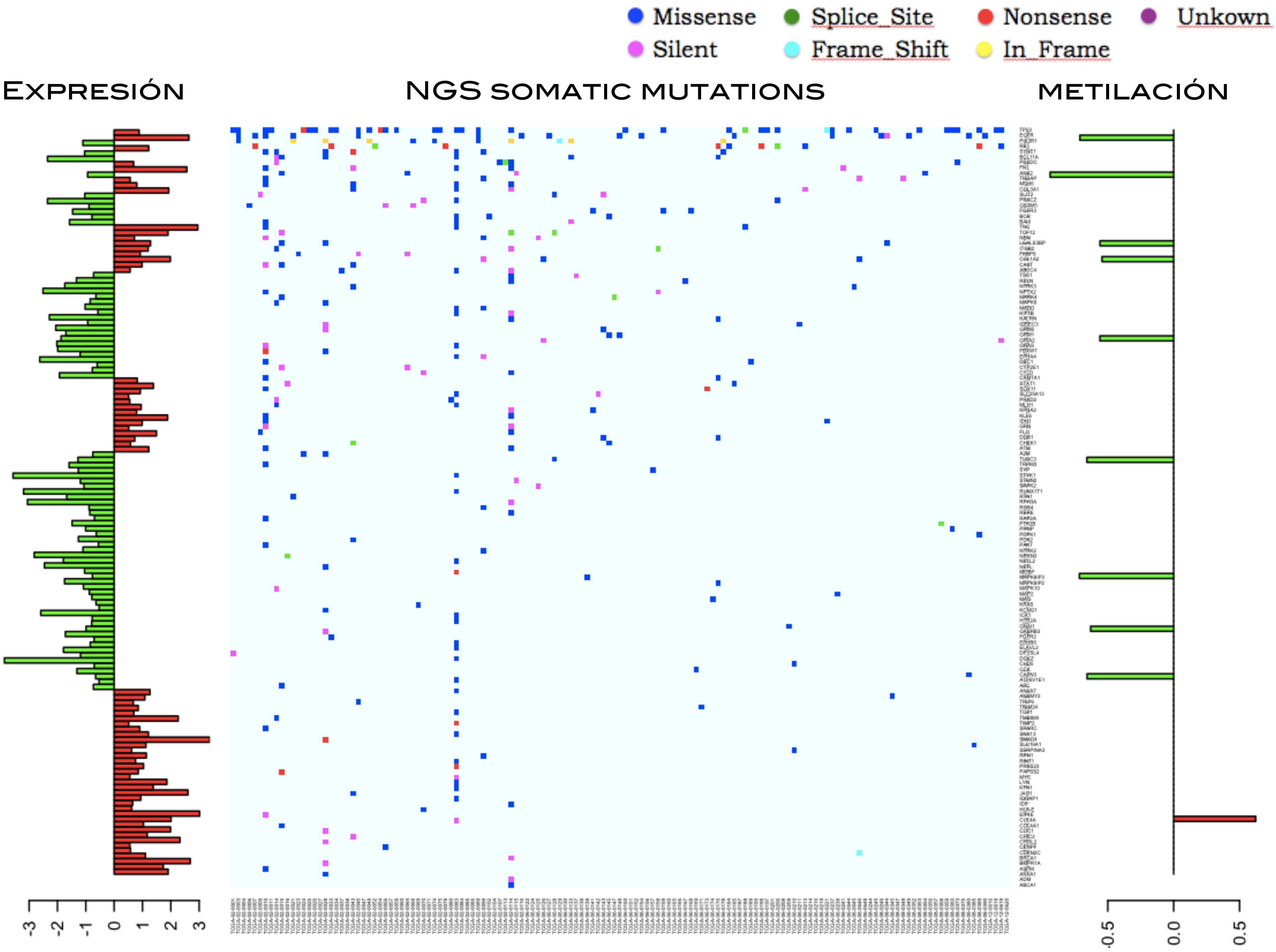
possible combinations (scenarios)

- How many do make sense?
- How many do we have?

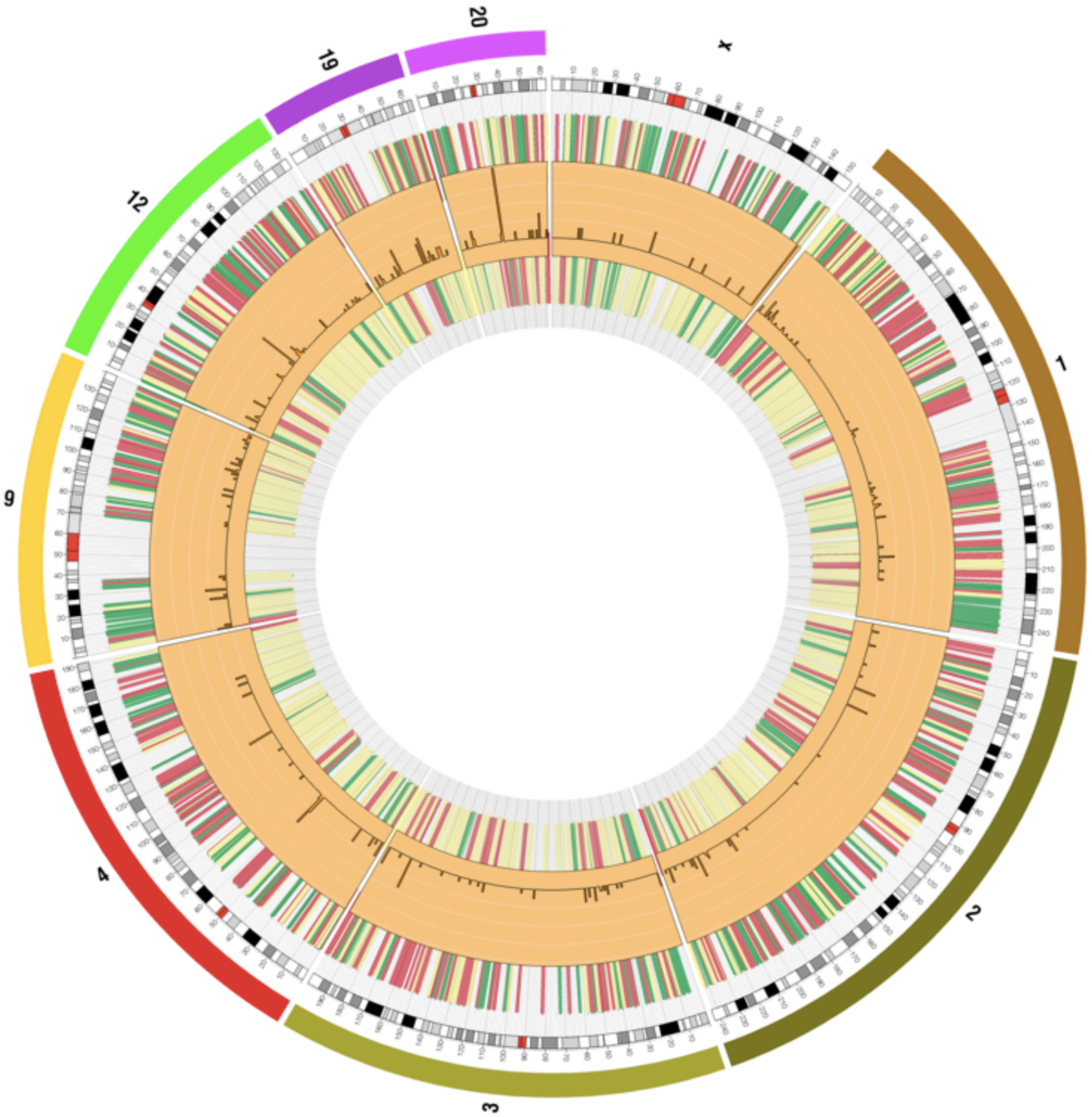
Results & Visualization

3-Platform Integration
2 Platforms for validation


{Mutations, Methylation, mRNA}
1400
1200
1000
800
600
400
200
0
49
328
28
1274
0
1418
31
475
28
6
9
3
71
331
62
0
14
1
{0-1-1}
{0-10}
{0-11}
{00-1}
{000}
{001}
{01-1}
{010}
{011}
{1-1-1}
{1-10}
{1-11}
{10-1}
{100}
{101}
{11-1}
{110}
{111}

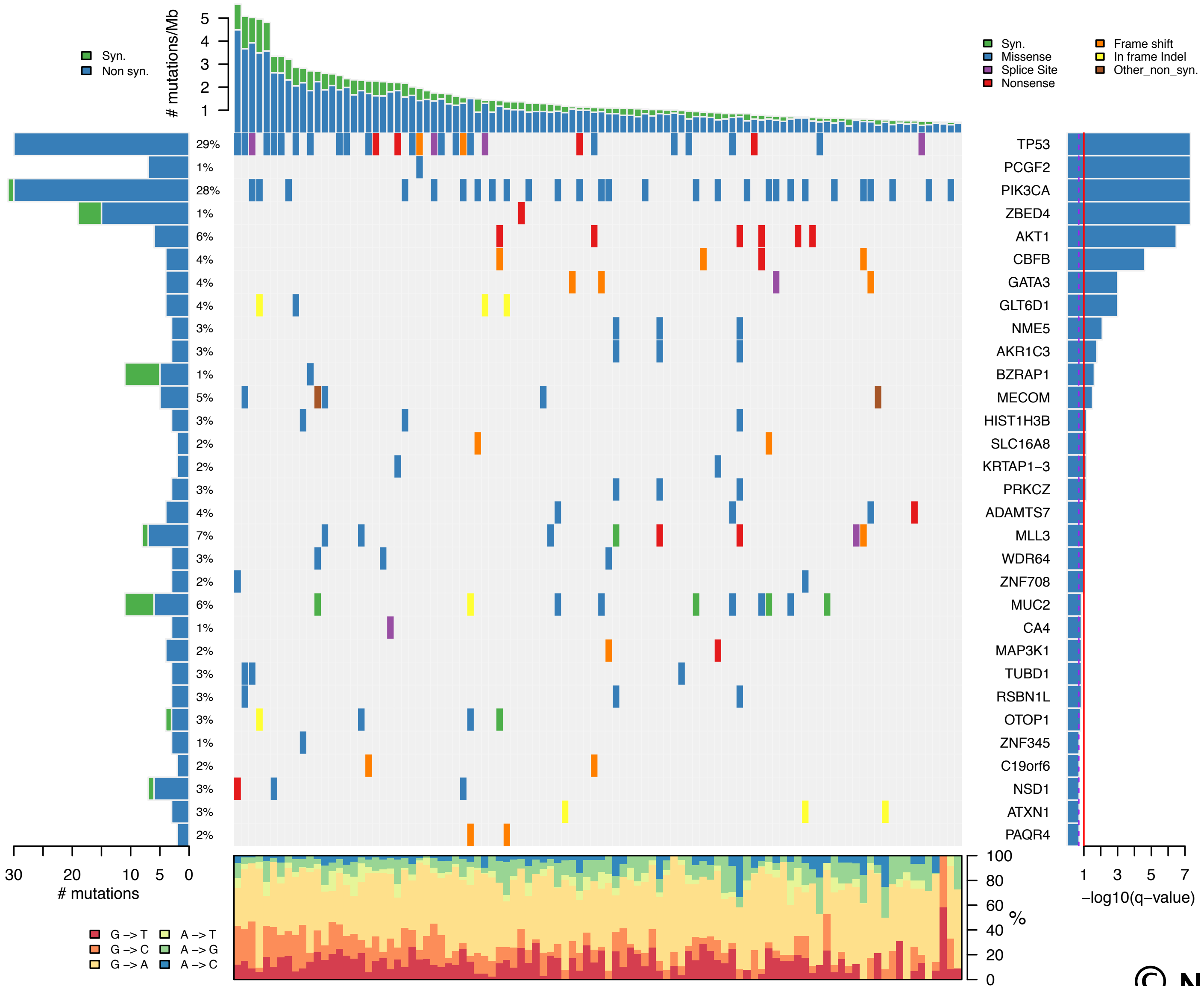

● Missense
● Splice_Site
● Nonsense
● Unkown
● Silent
● Frame_Shift
● In_Frame
EXPRESIÓN
NGS SOMATIC MUTATIONS
METILACIÓN
-3
-2
-1
0
1
2
3
-0.5
0.0
0.5

visualization



National Institute of Genomic Medicine MEXICO

visualization

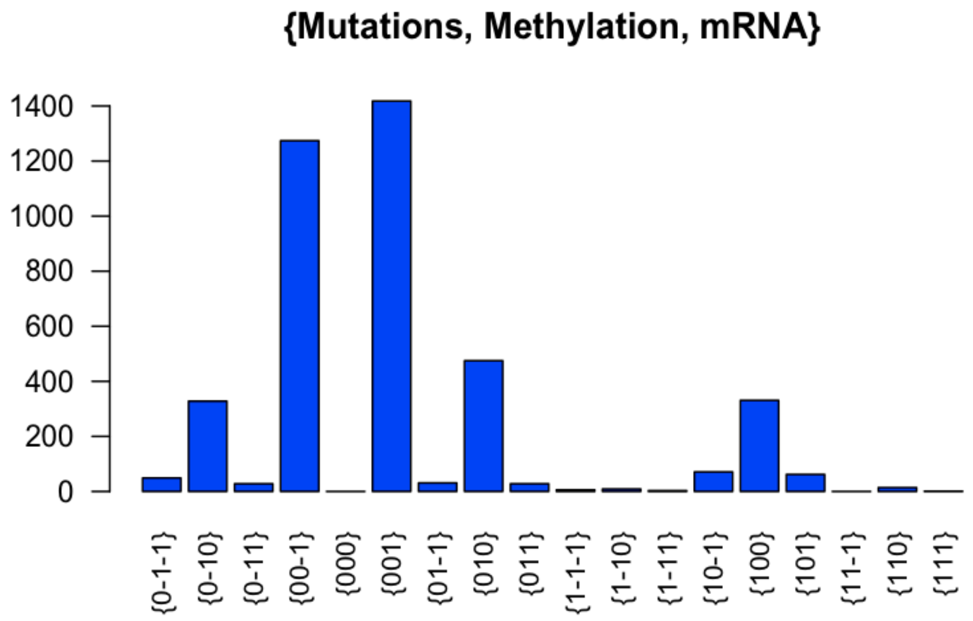


© Nicolas Stransky

Discussion & Remarks

A case of study
3-MDI as tool
Noise classification
CpG islands and methylation profiles
Machine Learning

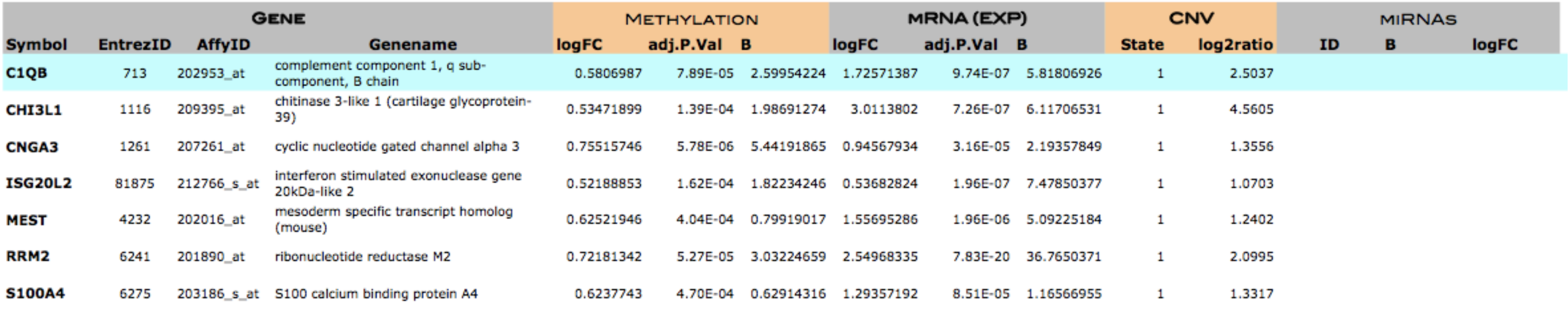
A case of study



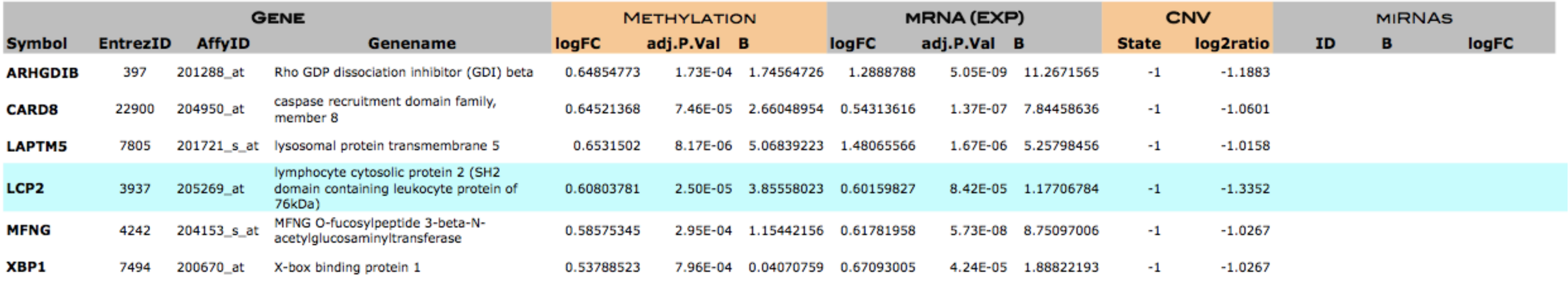
- No mutations present
- Hyper-methylated
- Up-regulated

{0, 1 , 1}

A case of study

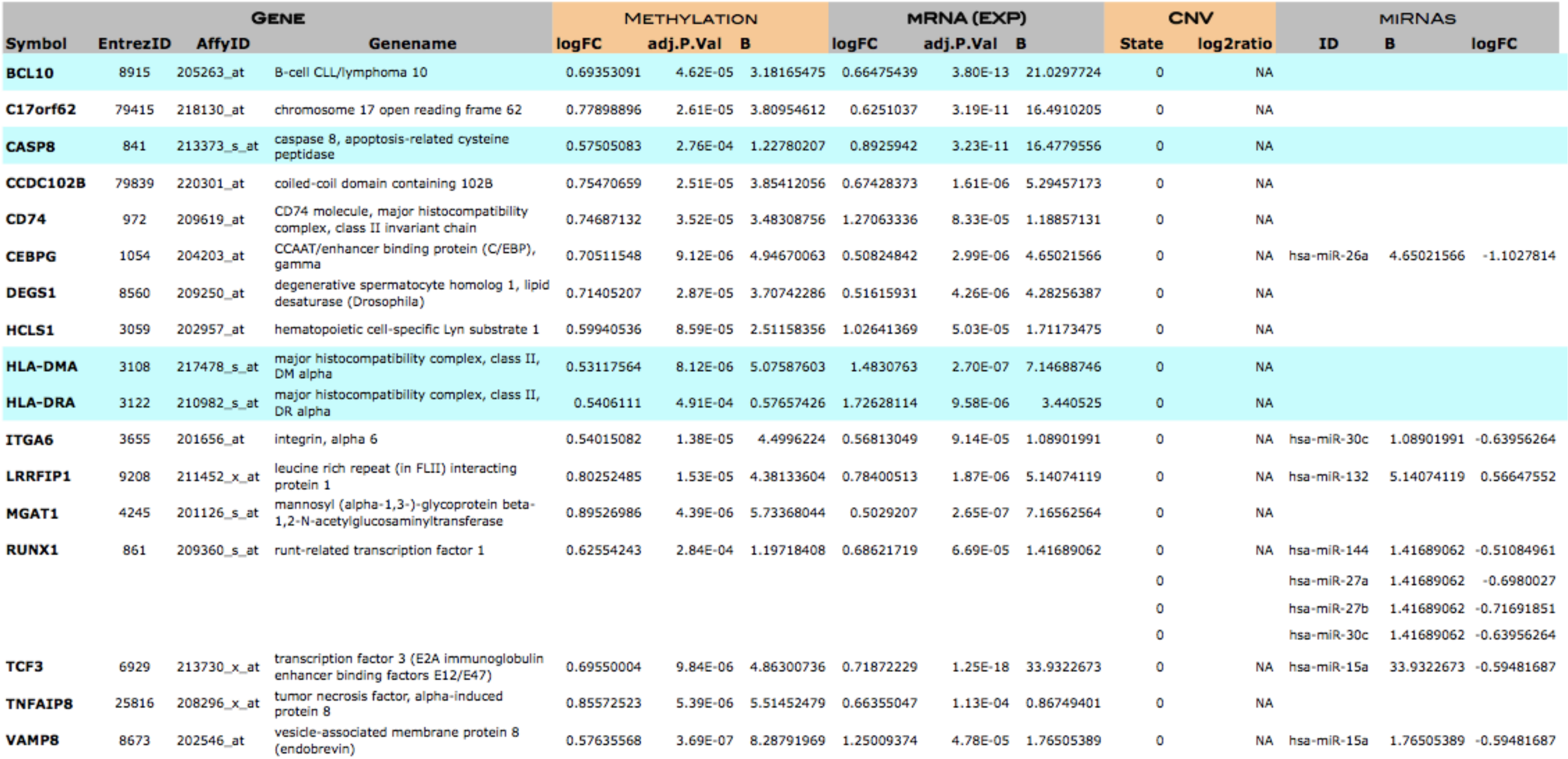


GENE				METHYLATION			MRNA (EXP)			CNV		MIRNAS		
Symbol	**EntrezID**	**AffyID**	**Genename**	**logFC**	**adj.P.Val**	**B**	**logFC**	**adj.P.Val**	**B**	**State**	**log2ratio**	**ID**	**B**	**logFC**
C1QB	713	202953_at	complement component 1, q sub-component, B chain	0.5806987	7.89E-05	2.59954224	1.72571387	9.74E-07	5.81806926	1	2.5037			
CHI3L1	1116	209395_at	chitinase 3-like 1 (cartilage glycoprotein-39)	0.53471899	1.39E-04	1.98691274	3.0113802	7.26E-07	6.11706531	1	4.5605			
CNGA3	1261	207261_at	cyclic nucleotide gated channel alpha 3	0.75515746	5.78E-06	5.44191865	0.94567934	3.16E-05	2.19357849	1	1.3556			
ISG20L2	81875	212766_s_at	interferon stimulated exonuclease gene 20kDa-like 2	0.52188853	1.62E-04	1.82234246	0.53682824	1.96E-07	7.47850377	1	1.0703			
MEST	4232	202016_at	mesoderm specific transcript homolog (mouse)	0.62521946	4.04E-04	0.79919017	1.55695286	1.96E-06	5.09225184	1	1.2402			
RRM2	6241	201890_at	ribonucleotide reductase M2	0.72181342	5.27E-05	3.03224659	2.54968335	7.83E-20	36.7650371	1	2.0995			
S100A4	6275	203186_s_at	S100 calcium binding protein A4	0.6237743	4.70E-04	0.62914316	1.29357192	8.51E-05	1.16566955	1	1.3317			



GENE				METHYLATION			MRNA (EXP)			CNV		MIRNAS		
Symbol	**EntrezID**	**AffyID**	**Genename**	**logFC**	**adj.P.Val**	**B**	**logFC**	**adj.P.Val**	**B**	**State**	**log2ratio**	**ID**	**B**	**logFC**
ARHGDIB	397	201288_at	Rho GDP dissociation inhibitor (GDI) beta	0.64854773	1.73E-04	1.74564726	1.2888788	5.05E-09	11.2671565	-1	-1.1883			
CARD8	22900	204950_at	caspase recruitment domain family, member 8	0.64521368	7.46E-05	2.66048954	0.54313616	1.37E-07	7.84458636	-1	-1.0601			
LAPTM5	7805	201721_s_at	lysosomal protein transmembrane 5	0.6531502	8.17E-06	5.06839223	1.48065566	1.67E-06	5.25798456	-1	-1.0158			
LCP2	3937	205269_at	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)	0.60803781	2.50E-05	3.85558023	0.60159827	8.42E-05	1.17706784	-1	-1.3352			
MFNG	4242	204153_s_at	MFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	0.58575345	2.95E-04	1.15442156	0.61781958	5.73E-08	8.75097006	-1	-1.0267			
XBP1	7494	200670_at	X-box binding protein 1	0.53788523	7.96E-04	0.04070759	0.67093005	4.24E-05	1.88822193	-1	-1.0267			

A case of study



GENE				METHYLATION			MRNA (EXP)			CNV		MIRNAS		
Symbol	EntrezID	AffyID	Genename	logFC	adj.P.Val	B	logFC	adj.P.Val	B	State	log2ratio	ID	B	logFC
BCL10	8915	205263_at	B-cell CLL/lymphoma 10	0.69353091	4.62E-05	3.18165475	0.66475439	3.80E-13	21.0297724	0	NA			
C17orf62	79415	218130_at	chromosome 17 open reading frame 62	0.77898896	2.61E-05	3.80954612	0.6251037	3.19E-11	16.4910205	0	NA			
CASP8	841	213373_s_at	caspase 8, apoptosis-related cysteine peptidase	0.57505083	2.76E-04	1.22780207	0.8925942	3.23E-11	16.4779556	0	NA			
CCDC102B	79839	220301_at	coiled-coil domain containing 102B	0.75470659	2.51E-05	3.85412056	0.67428373	1.61E-06	5.29457173	0	NA			
CD74	972	209619_at	CD74 molecule, major histocompatibility complex, class II invariant chain	0.74687132	3.52E-05	3.48308756	1.27063336	8.33E-05	1.18857131	0	NA			
CEBPG	1054	204203_at	CCAAT/enhancer binding protein (C/EBP), gamma	0.70511548	9.12E-06	4.94670063	0.50824842	2.99E-06	4.65021566	0	NA	hsa-miR-26a	4.65021566	-1.1027814
DEGS1	8560	209250_at	degenerative spermatocyte homolog 1, lipid desaturase (Drosophila)	0.71405207	2.87E-05	3.70742286	0.51615931	4.26E-06	4.28256387	0	NA			
HCLS1	3059	202957_at	hematopoietic cell-specific Lyn substrate 1	0.59940536	8.59E-05	2.51158356	1.02641369	5.03E-05	1.71173475	0	NA			
HLA-DMA	3108	217478_s_at	major histocompatibility complex, class II, DM alpha	0.53117564	8.12E-06	5.07587603	1.4830763	2.70E-07	7.14688746	0	NA			
HLA-DRA	3122	210982_s_at	major histocompatibility complex, class II, DR alpha	0.5406111	4.91E-04	0.57657426	1.72628114	9.58E-06	3.440525	0	NA			
ITGA6	3655	201656_at	integrin, alpha 6	0.54015082	1.38E-05	4.4996224	0.56813049	9.14E-05	1.08901991	0	NA	hsa-miR-30c	1.08901991	-0.63956264
LRRFIP1	9208	211452_x_at	leucine rich repeat (in FLII) interacting protein 1	0.80252485	1.53E-05	4.38133604	0.78400513	1.87E-06	5.14074119	0	NA	hsa-miR-132	5.14074119	0.56647552
MGAT1	4245	201126_s_at	mannosyl (alpha-1,3-)-glycoprotein beta-1,2-N-acetylglucosaminyltransferase	0.89526986	4.39E-06	5.73368044	0.5029207	2.65E-07	7.16562564	0	NA			
RUNX1	861	209360_s_at	runt-related transcription factor 1	0.62554243	2.84E-04	1.19718408	0.68621719	6.69E-05	1.41689062	0	NA	hsa-miR-144	1.41689062	-0.51084961
										0		hsa-miR-27a	1.41689062	-0.6980027
										0		hsa-miR-27b	1.41689062	-0.71691851
										0		hsa-miR-30c	1.41689062	-0.63956264
TCF3	6929	213730_x_at	transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)	0.69550004	9.84E-06	4.86300736	0.71872229	1.25E-18	33.9322673	0	NA	hsa-miR-15a	33.9322673	-0.59481687
TNFAIP8	25816	208296_x_at	tumor necrosis factor, alpha-induced protein 8	0.85572523	5.39E-06	5.51452479	0.66355047	1.13E-04	0.86749401	0	NA			
VAMP8	8673	202546_at	vesicle-associated membrane protein 8 (endobrevin)	0.57635568	3.69E-07	8.28791969	1.25009374	4.78E-05	1.76505389	0	NA	hsa-miR-15a	1.76505389	-0.59481687

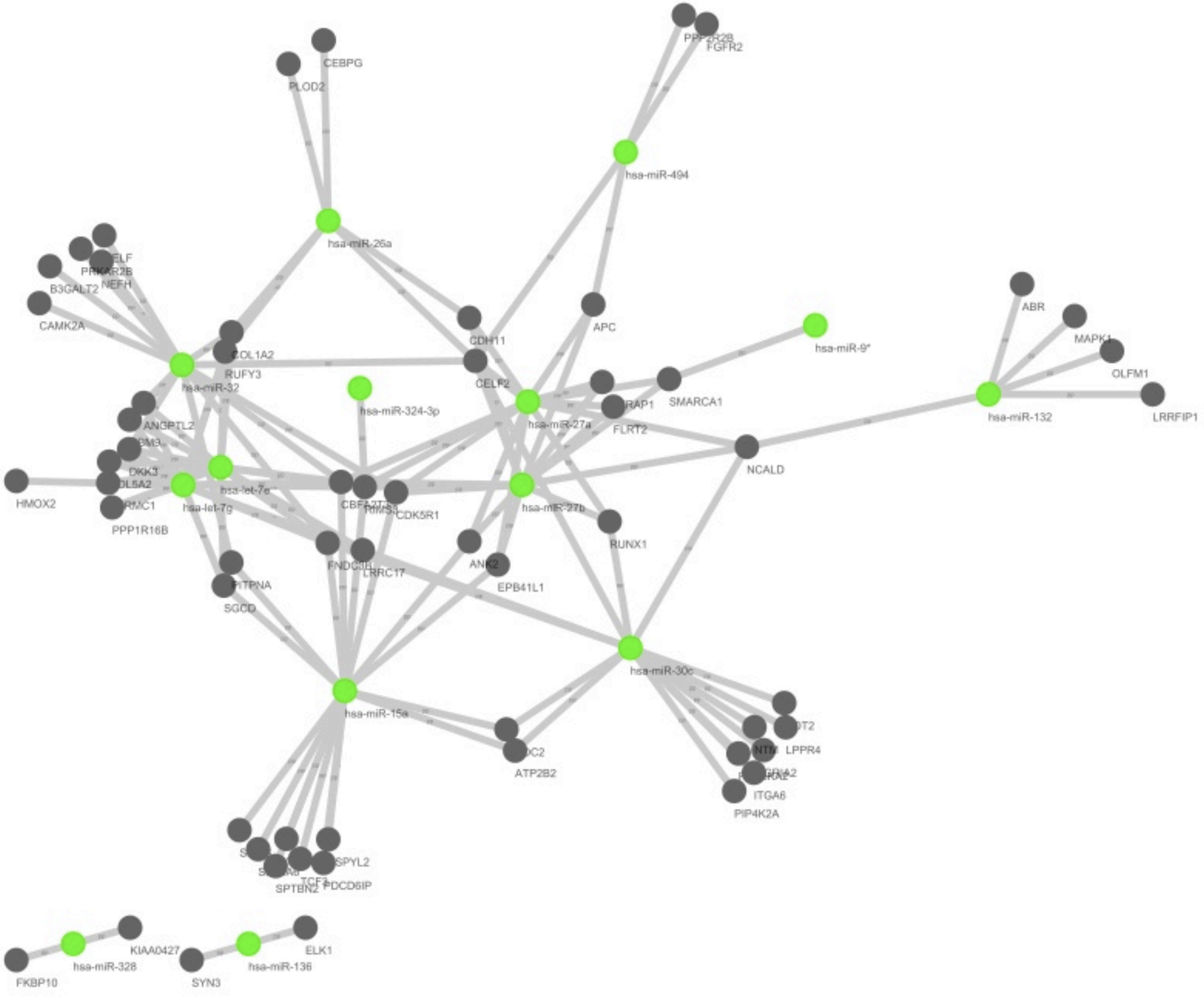
DNA methylation—miRNA network analysis

- Integrated analysis of DNA methylation profiles in CpG islands and miRNA differential expression can be explored with our 3-state model

- It can also be represented in a network-based analysis

- Results suggest that DNA methylation and miRNA transcriptional regulation are closely related for a particular state-vector representing a novel characteristic pattern.

DNA methylation—miRNA network analysis



DNA methylation—miRNA network analysis

- For instance, this analysis shows 9 miRNAs related (as putative targets) to genes over-represented with respect to changes in the CpG methylation status.

- That is, genes whose methylation profiles and miRNA targeting status may potentially affect their corresponding mRNA expression levels.

- Pathway enrichment analysis using GO for this set of 9 genes shows only a few pathways significantly enriched in biological processes mainly involved in neuronal functions (eg. Axon guidance, synaptical transmission)

Remarks

- Data driven approaches for large multiplatform data may fit better than biological ones
- Each additional genomic dimension increases both the amount of information and consequently the biological and computational complexity of the analysis
- Noise behavior should be explored
- Machine learning approaches can be applied regardless of the number of platforms
- Bayesian approach might be improved by the prior information from the counts