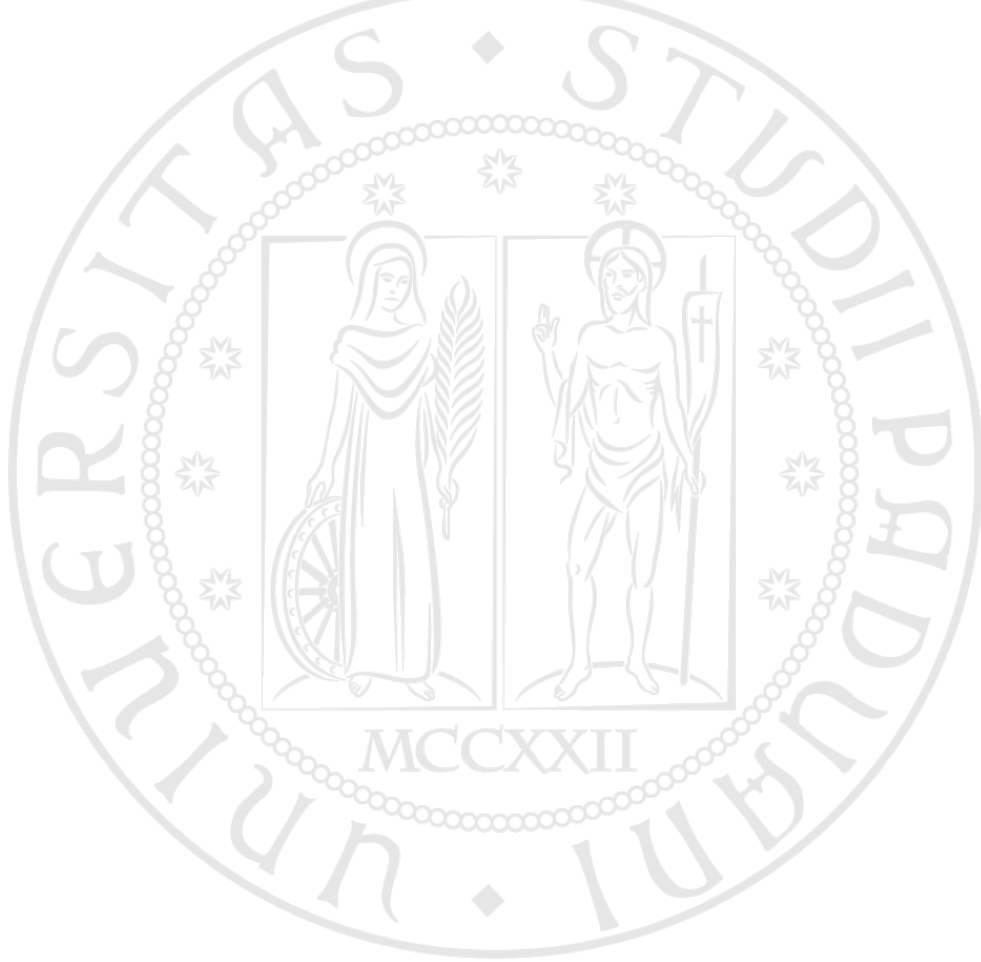


CAGI slide pack guidelines

I agree to my slides being used according to the following terms: [Marco Carraro](#)

- Authors; please include THIS slide at the beginning of your presentation to indicate that you agree with reuse of your slides in CAGI presentations, according to the terms below, and digitally sign by writing your name above.
- To disseminate the CAGI results, I encourage all CAGI participants to remix these CAGI slides for inclusion in their oral and poster presentations
- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 2013. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.





MARCO CARRARO

*BioComputing UP, Dept. of Biomedical Sciences,
University of Padova Italy*

PCM1: predict the effects of missense changes in a zebrafish model

CAGI

05/07/2018

Overview

- PCMI1 functions and characteristics
- The challenge
- Dataset
- Assessment metrics
- Results

Data providers

Prof. Nicholas Katsanis

Director of Center for Human Disease Modeling
Duke University



Maria Kousi, PhD

MIT Computer Science and Artificial Intelligence Laboratory
(CSAIL)
Broad Institute of MIT and Harvard



Pericentriolar Material 1 (PCM1):

predict the effects of missense changes in a zebrafish model

Schizophrenia

Schizophrenia is a severe psychiatric disorder

- positive symptoms
- negative symptoms
- cognitive impairment

20% of patients experience chronic symptoms

Life expectancy is reduced by 10-20 years

Schizophrenia	
Core features	Delusions, hallucinations, impaired motivation, reduction in spontaneous speech, and social withdrawal
Additional features	Cognitive impairment, episodes of elated mood, and episodes of depressive mood
Heritability estimate	Roughly 80%
Average age of onset	16–30 years
Drug treatment	Antipsychotics

Mod. from Owen et al., Lancet, 2016

Genetic factors contribute substantially, but not exclusively the onset

highly polygenic – GWAS identified >100 genetic loci

rare SNPs, indels and CNVs

common risk variants with:

- bipolar disorder
- major depressive disorder
- autism spectrum disorder (ASD)

Overlapping mechanisms between these disorders have been hypothesized

Lee et al., Nature Genetics, 2013

PCM1

Several studies have implicated PCM1 variants as risk factors for schizophrenia. *Kamiya et al., Arch Gen Psychiatry, 2018*

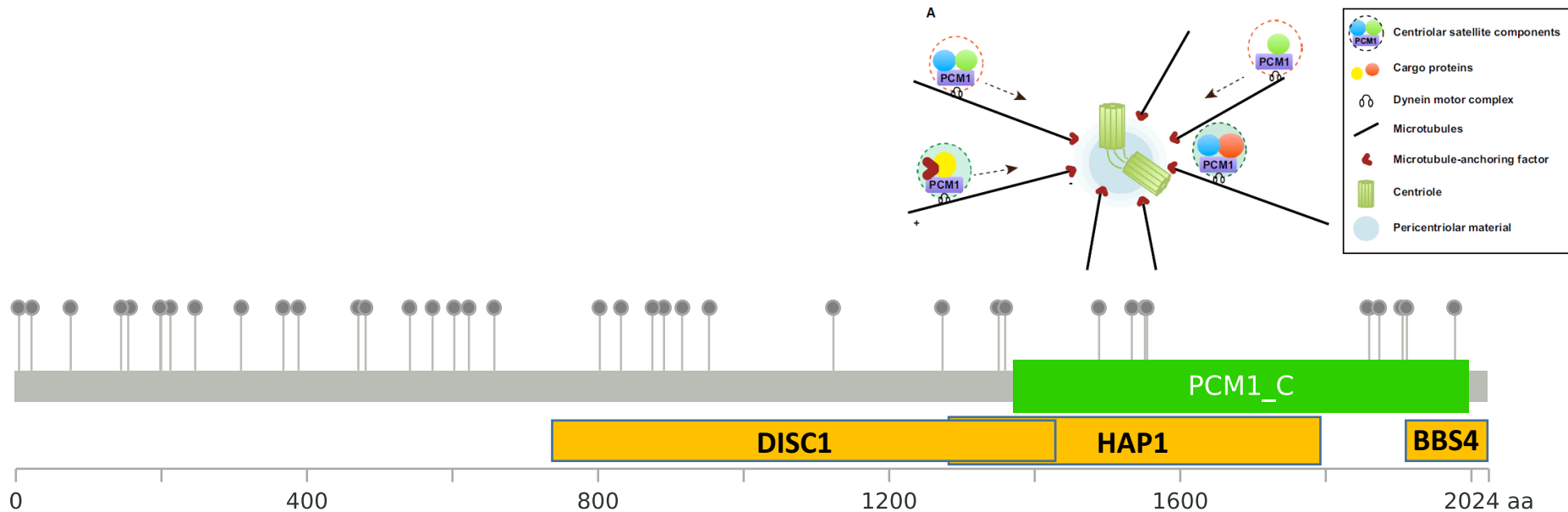
Down regulation of PCM1 repress neuronal proliferation and differentiation. *Zhang et al., Elife, 2010*

PCM1 mutant zebrafish show brain ventricle alterations.

The canonical transcript of PCM1 encodes a protein of 2024 amino acids.

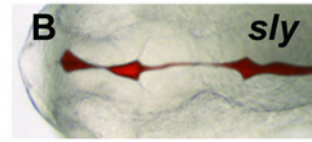
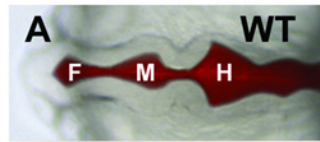
The protein contains coiled coil regions between areas of low complexity.

More than 100 interactors have been identified. *Hori and Toda, Cell Mol Life Sci., 2010*



mod. from Kamiya et al., Arch Gen Psychiatry, 2018

PCM1 mutants in zebrafish



mod. from Lowery et al., Anat Rec, 2010

Katsanis lab injected a group of embryos with MO+mRNA of the human gene carrying a mutation.

MO+VAR vs MO

p-value

MO+VAR vs MO+WT

p-value

MO+VAR $\neq$ MO $\cup$ MO+VAR $\neq^*$ MO+WT $\rightarrow$ Pathogenic / LOF

MO+VAR $\neq^*$ MO $\cup$ MO+VAR $\neq$ MO+WT $\rightarrow$ Benign

MO+VAR $\neq^*$ MO $\cup$ MO+VAR $\neq^*$ MO+WT $\rightarrow$ Hypomorphic / Partially LOF

MO = morpholino

Experimental data

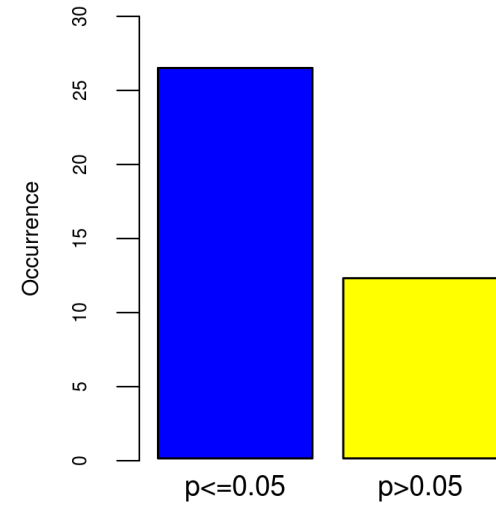
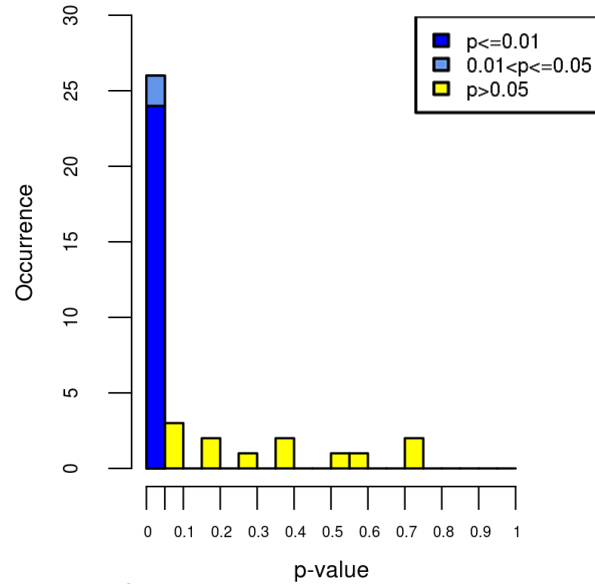
Variant	<i>p-value</i> MO+Var VS MO	<i>p-value</i> from MO+Var VS MO+WT	Functional effect (class)	Functional effect (description)
G6D	0.067	< 0.0001	2	Loss of function
E23D	0.0004	0.0007	1	Hypomorph
T77A	0.57	< 0.0001	2	Loss of function
M146V	< 0.0001	0.13	0	Benign
A156V	< 0.0001	0.0099	1	Hypomorph
M200T	0.28	< 0.0001	2	Loss of function
M200I	...	...	...	...

The challenge:

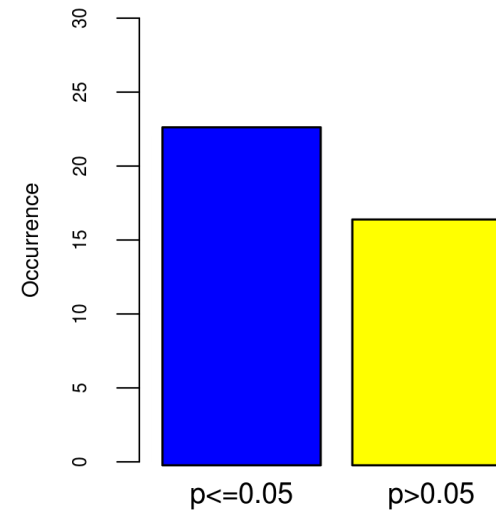
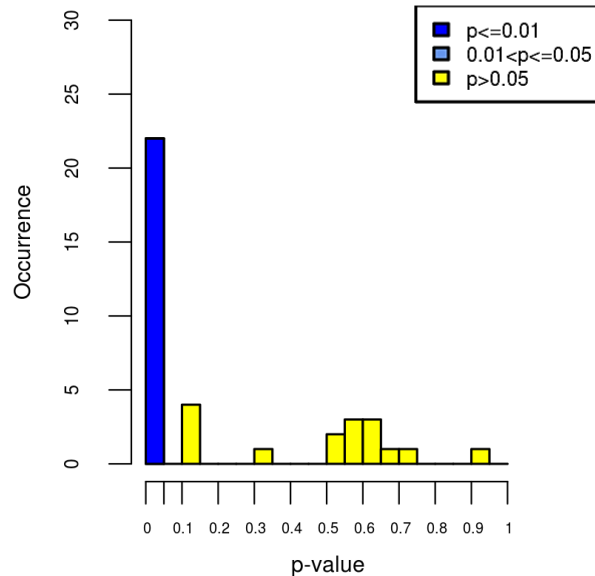
- predict ***p-value*** that **variant** is statistically different from **MO** (+ confidence)
- predict ***p-value*** that **variant** is statistically different from **MO+WT** (+ confidence)
- predict **functional effect** 0,1,2 (+ confidence)

Dataset

**p-values
MO+VAR vs MO**



**p-values
MO+VAR vs MO+WT**



Disclaimer

***p-values* have not been considered to rank predictions**

- which performance indices are suitable ?
- classification/regression?
- classification -> biological meaning of the classes?
- regression -> but we are assessing *p-values*?
- **no agreement among *p-values* and final predicted classes**



Variant	<i>p-value</i> MO+Var VS MO	<i>p-value</i> from MO+Var VS MO+WT	Functional effect (class)
G6D	0.067	< 0.0001	2
E23D	0.0004	0.0007	1
T77A	< 0.0001	0.55	2

Final assessment is based on final predicted classes only

Submissions

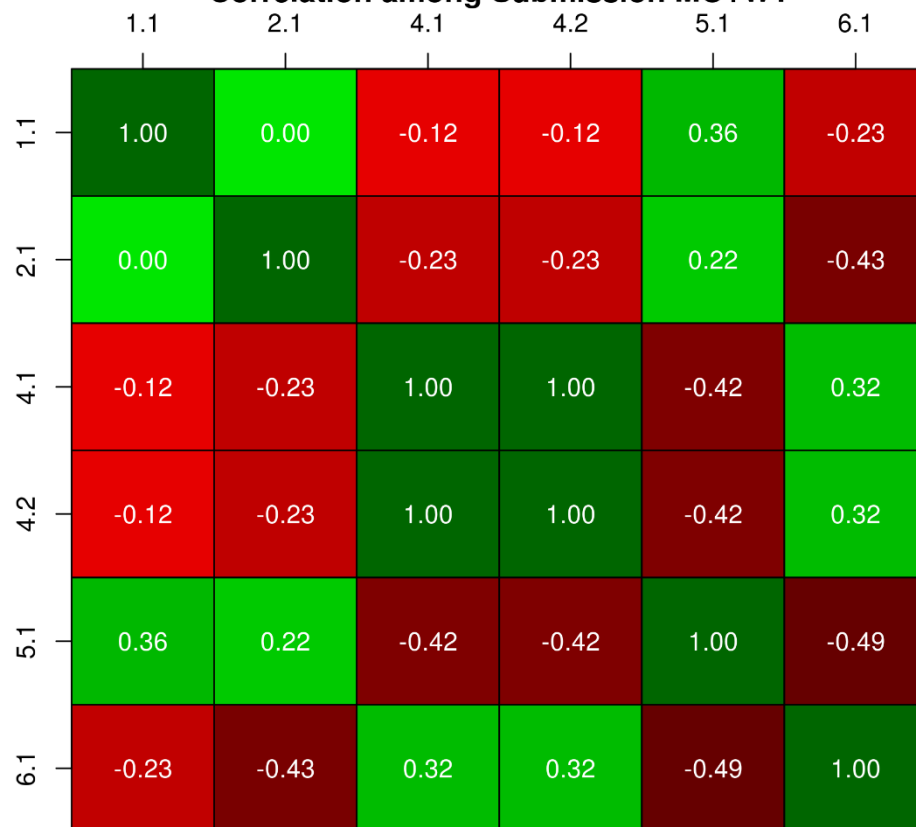
- 6 participating groups
- 7 submissions
- 1 empty submission

Group 1	BolognaBiocomputing
Group 2	Lichtarge Lab.
Group 4	Bromberg Lab.
Group 5	UCSD Cancer Bioinformatics Group
Group 6	BioFold

Correlation among Submission MO



Correlation among Submission MO+WT

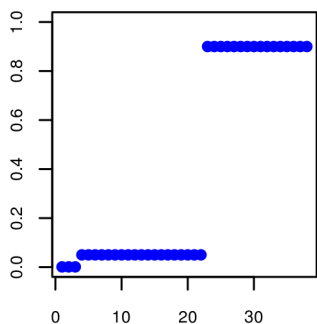


Submissions

p-values

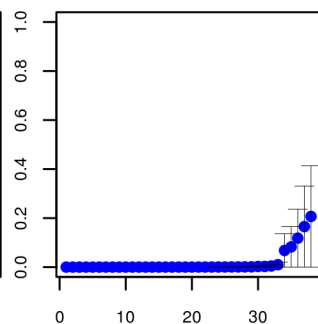
MO+VAR vs MO

Submission 1.1



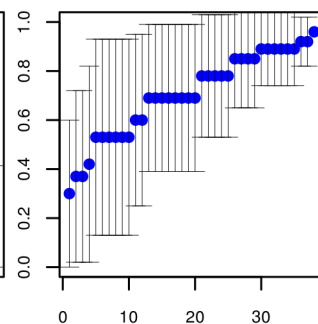
Predictions

Submission 2.1



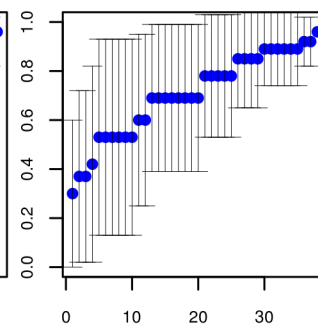
Predictions

Submission 4.1



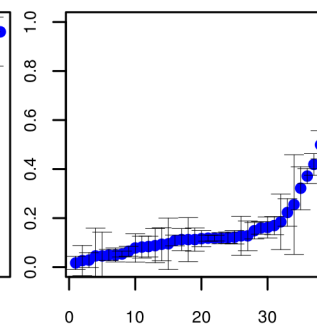
Predictions

Submission 4.2



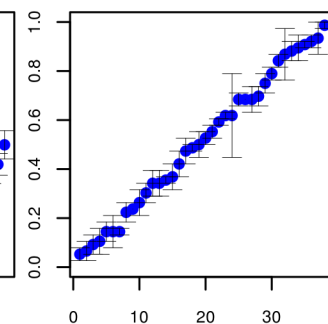
Predictions

Submission 5.1



Predictions

Submission 6.1

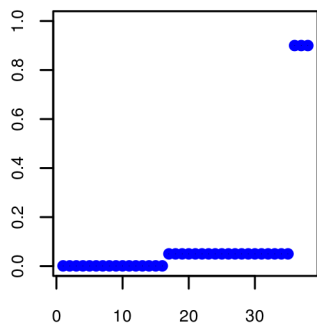


Predictions

p-values

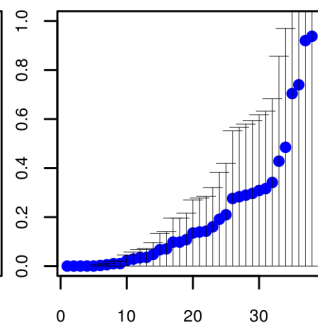
MO+VAR vs MO+WT

Submission 1.1



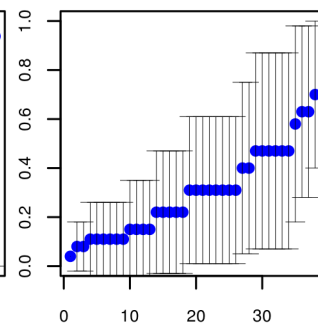
Predictions

Submission 2.1



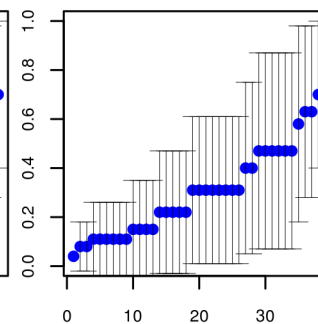
Predictions

Submission 4.1



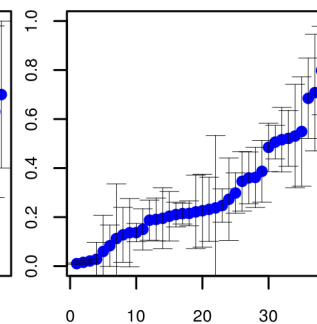
Predictions

Submission 4.2



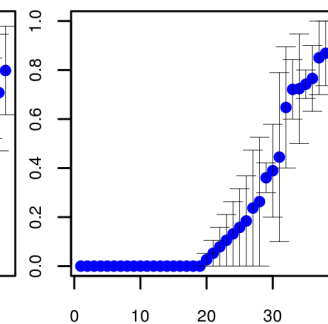
Predictions

Submission 5.1



Predictions

Submission 6.1



Predictions

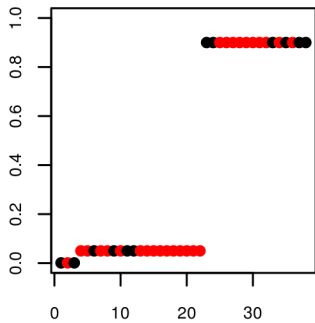
Classes vs Predicted *p*-values

p-values

MO+VAR vs MO

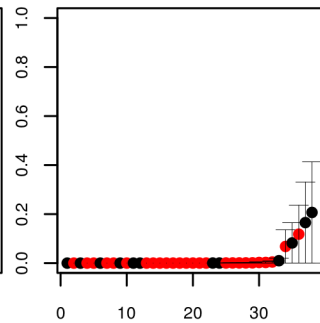
● $\geq 0,05$
● $< 0,05$

Submission 1.1



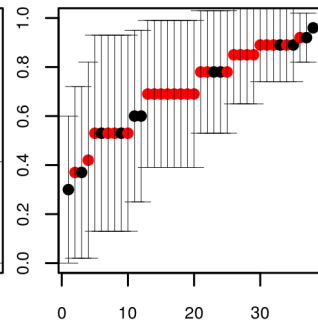
Predictions

Submission 2.1



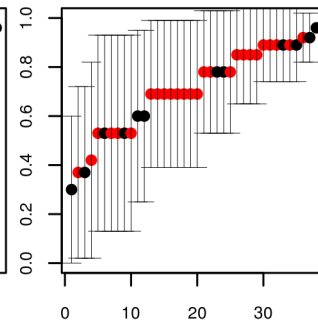
Predictions

Submission 4.1



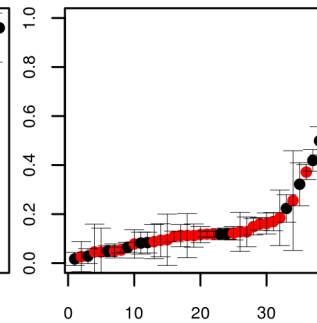
Predictions

Submission 4.2



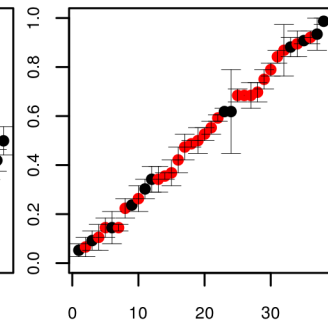
Predictions

Submission 5.1



Predictions

Submission 6.1

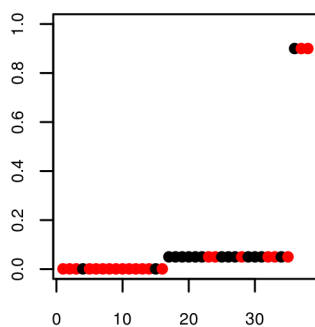


Predictions

p-values

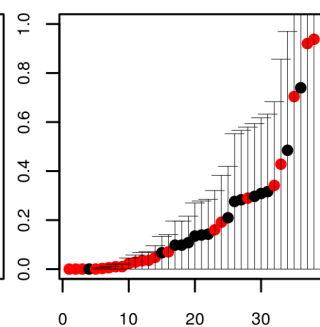
MO+VAR vs MO+WT

Submission 1.1



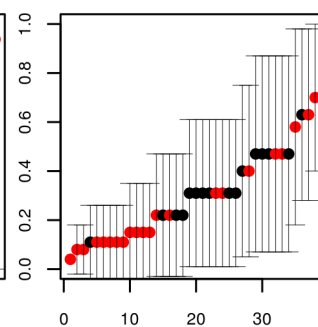
Predictions

Submission 2.1



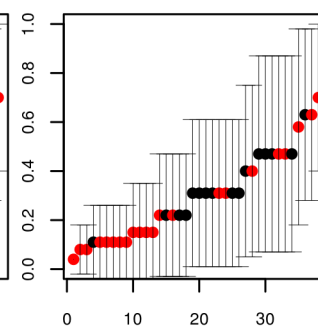
Predictions

Submission 4.1



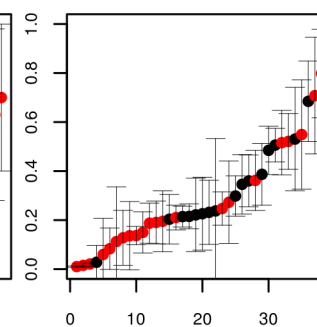
Predictions

Submission 4.2



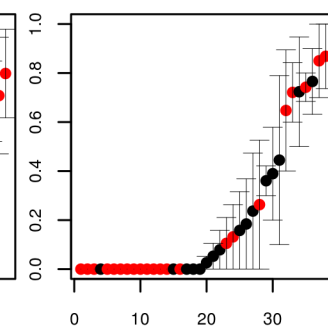
Predictions

Submission 5.1



Predictions

Submission 6.1



Predictions

Performance assessment

	BACC	MCC	F1	TPR	TNR
Submission 4.1	0.66	0.35	0.59	0.45	0.88
Submission 6.1	0.54	0.08	0.53	0.45	0.62
Submission 1.1	0.43	-0.25	0.67	0.86	0.00
Submission 2.1	0.49	-0.01	0.44	0.36	0.62
Submission 4.2	0.49	-0.04	0.08	0.05	0.94
Submission 5.1	0.34	-0.35	0.50	0.55	0.12

BACC = Balanced Accuracy

MCC = Matthews Correlation Coefficient

F1 = F1 score

TPR = True Positive Rate

TNR = True Negative Rate

Performance assessment

	BACC	MCC	F ₁	TPR	TNR	AVG	FINAL
Submission 4.1	1 (0.67)	1 (0.35)	2 (0.59)	3.5 (0.46)	2 (0.88)	1.9	1
Submission 6.1	2 (0.54)	2 (0.08)	3 (0.53)	3.5 (0.46)	3.5 (0.63)	2.8	2
Submission 1.1	5 (0.43)	5 (-0.25)	1 (0.67)	1 (0.86)	6 (0)	3.6	3
Submission 2.1	3 (0.49)	3 (-0.01)	5 (0.44)	5 (0.36)	3.5 (0.63)	3.9	4
Submission 4.2	4 (0.49)	4 (-0.04)	6 (0.08)	6 (0.05)	1 (0.94)	4.2	5
Submission 5.1	6 (0.34)	6 (-0.35)	4 (0.5)	2 (0.55)	5 (0.13)	4.6	6

BACC = Balanced Accuracy

MCC = Matthews Correlation Coefficient

F1 = F1 score

TPR = True Positive Rate

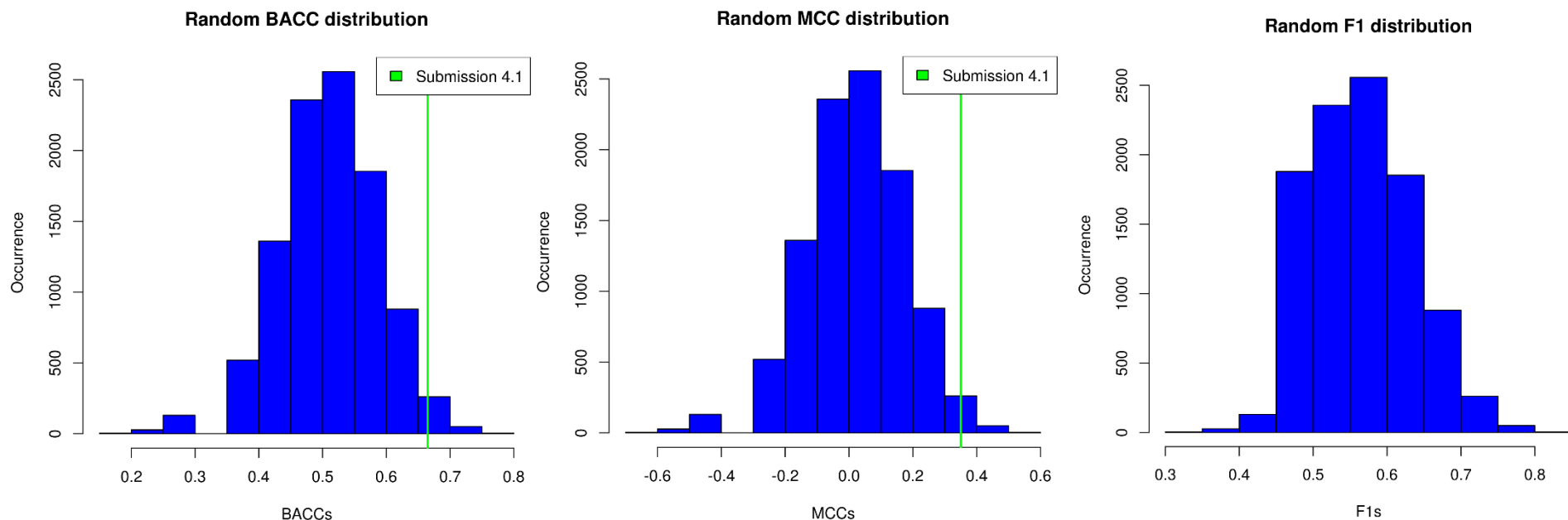
TNR = True Negative Rate

Confusion matrices

Submission 4.1			Submission 6.1			Submission 1.1		
	Obs Others	Obs Benign		Obs Others	Obs Benign		Obs Others	Obs Benign
Pred Others	10	2	Pred Others	10	6	Pred Others	19	16
Pred Benign	12	14	Pred Benign	12	10	Pred Benign	3	0
PPV	0.833		PPV	0.625		PPV	0.543	
NPV	0.539		NPV	0.455		NPV	0	
TPR	0.455		TPR	0.455		TPR	0.864	
TNR	0.875		TNR	0.625		TNR	0	
MCC	0.350		MCC	0.080		MCC	-0.250	
ACC	0.632		ACC	0.526		ACC	0.5	
BACC	0.665		BACC	0.540		BACC	0.432	
F1	0.588		F1	0.526		F1	0.667	
Submission 2.1			Submission 4.2			Submission 5.1		
	Obs Others	Obs Benign		Obs Others	Obs Benign		Obs Others	Obs Benign
Pred Others	8	6	Pred Others	1	1	Pred Others	12	14
Pred Benign	14	10	Pred Benign	21	15	Pred Benign	10	2
PPV	0.571		PPV	0.5		PPV	0.462	
NPV	0.417		NPV	0.417		NPV	0.167	
TPR	0.363		TPR	0.046		TPR	0.546	
TNR	0.625		TNR	0.935		TNR	0.125	
MCC	-0.011		MCC	-0.038		MCC	-0.350	
ACC	0.474		ACC	0.421		ACC	0.368	
BACC	0.494		BACC	0.492		BACC	0.335	
F1	0.444		F1	0.083		F1	0.5	

Statistical significance

- 10000 random predictions
- same bias between classes as real data



Comparison with State of the art

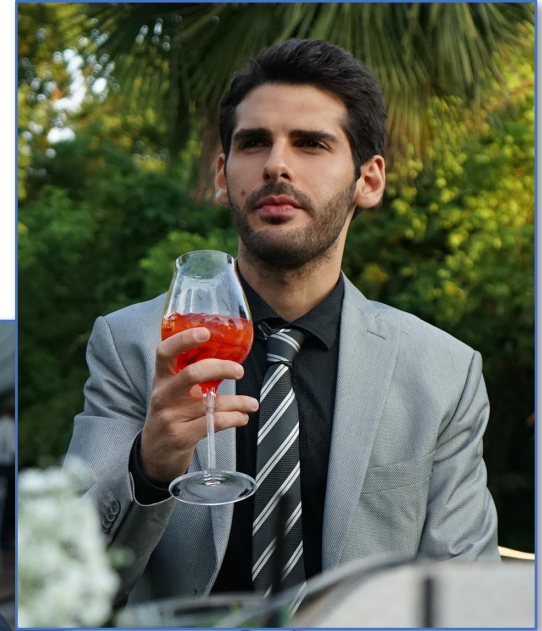
PHD-SNP
SNAP
SNPs&GO
EvolutionaryAction
VEST

Acknowledgments

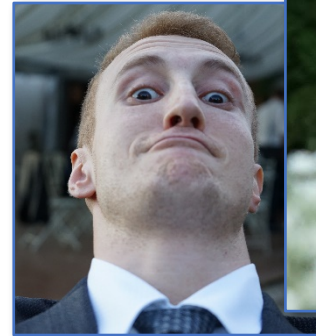
Prof. Silvio CE Tosatto



Luigi Chiricosta MSc



Marco Necci
PhD Candidate



**CAGI Organizers
Data providers
Predictors**

Francesco Reggiani
PhD Candidate



CAGI

