

CAGI slide pack guidelines

I agree to my slides being used according to the following terms:

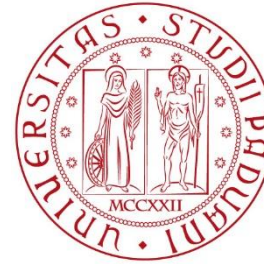
Emanuela Leonardi

- 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 6. 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.



Predict patients' clinical descriptions and pathogenic variants from gene panel sequences

**Laboratory of Molecular Genetics of
Neurodevelopmental Disorders (NDDs)**
Department of Woman and Children Health
University of Padova, Italy
Centro di Ricerca Pediatrica/Città della
Speranza



Providers



Prof. Alessandra
Murgia, MD PHD



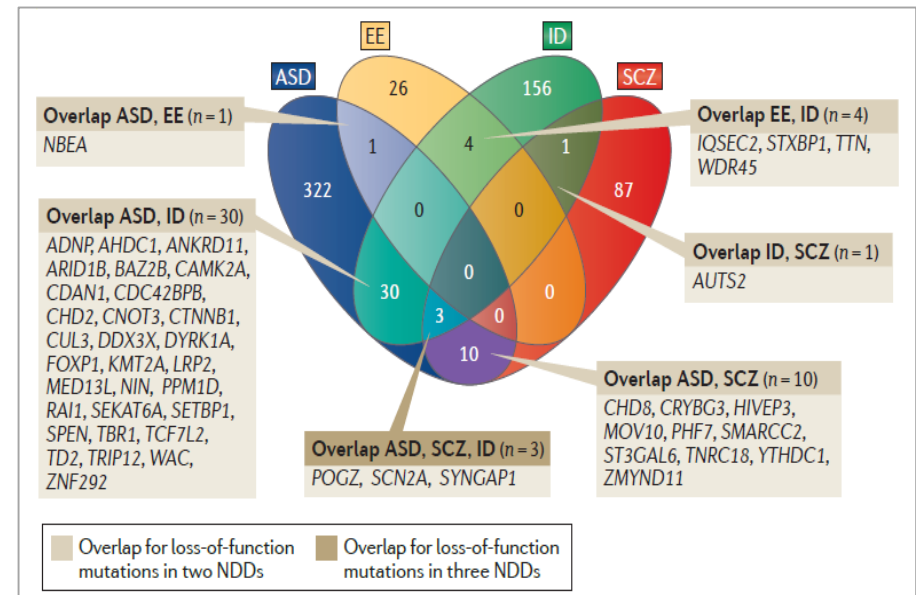
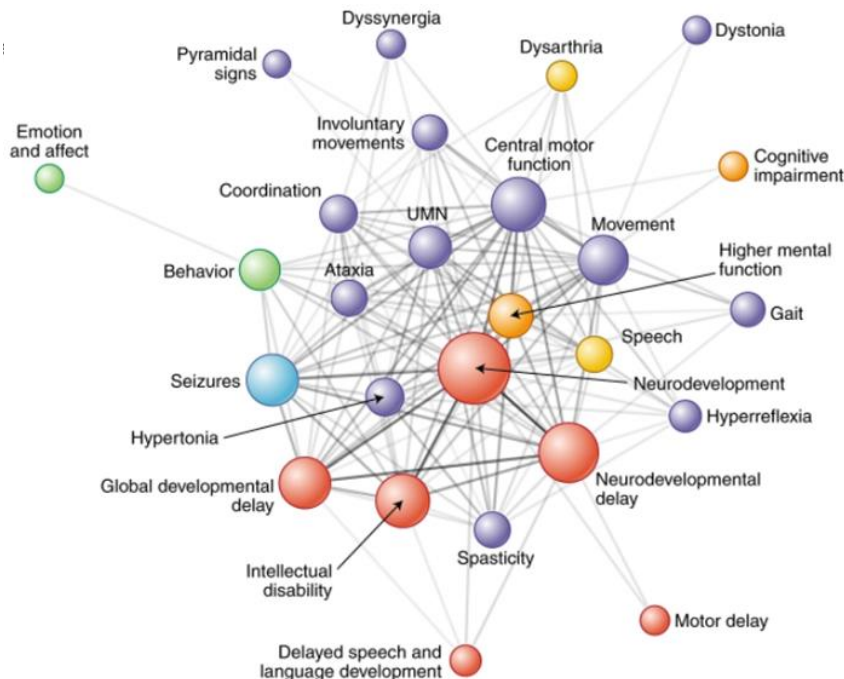
Dr. Emanuela
Leonardi, PhD



Dr. Maria Cristina
Aspromonte, PhD

Neurodevelopmental disorders (NDDs)

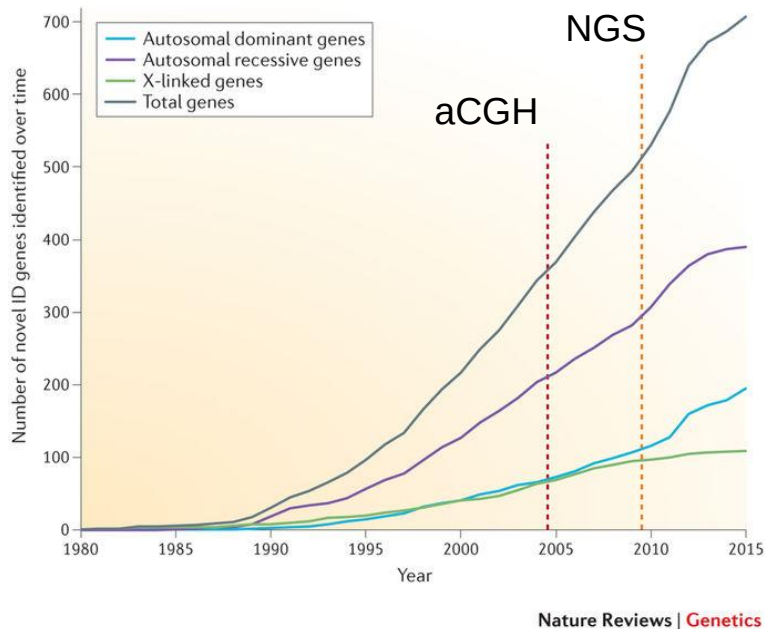
NDDs are characterized by significant comorbidities such as intellectual disability (ID), Autism Spectrum Disorders (ASD), and epilepsy.



Genes associated with different NDDs are shared among these disorders and converge on common functional pathways

Sanders 2019 Nat Med

Neurodevelopmental disorders (NDDs)



Vissers LELM et al., Nature reviews Genetics, 2016

✓ Wide genotype heterogeneity

18% of protein-coding genes implicated in NDDs

✓ Wide phenotypic variability

Phenotype knowledge biased towards severely affected cases, with milder cases and atypical presentations underrepresented

✓ Complex genetic architecture

Sporadic cases of ASD/ID are enriched for disruptive *de novo* mutations

Contribution of rare and common variants

Custom ID-gene panel: how it was designed

- ✓ Genes associated with **non-syndromic NDDs** characterized by ID and/or ASD
 - Dominant genes (44)
 - Recessive genes (8)
 - AD/AR genes (3)
 - XLID genes (19)
- ✓ The **74 panel genes** are more connected in an interaction network than a set of randomly selected genes.
- ✓ **22 genes** associated with Rett-like spectrum disorders
- ✓ **10 genes candidate for ASD**, with single piece of evidence in the literature.
- ✓ Low penetrance



Aspromonte et al., Hum Mut 2019

ID-gene panel: sequencing workflow

The ION Torrent Platform Semiconductor sequencing technology

increasingly applied in clinical and
diagnostic settings:

- low cost
- rapid turnaround time (48h)
- flexibility



Target enrichment by multiplex-PCR amplification

- Reduced time for library preparation
- High Coverage
- High accuracy at SNPs discovery
- **High ratio of false positive indels** (multi-allelic variant calls 12% FP, recurrent FP in homopolymeric regions and triplet repeats)

1834 amplicons (520Kb)

415 cases from 76 runs (1018 to 2021)

97% reads on targets

Mean Depth of Coverage >100x (range 100x-1000x)

ID-gene panel: data analysis and interpretation

~ **300 variants** called for
each VCF file

~ **10 variants filtered** for each
VCF file

~ **2/3 variants selected** for each
patient

Filtering

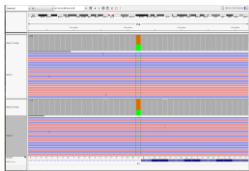
Remove SNVs with MAF
>0.0002 (AD genes)



Select SNVs with frequency
<5/N tested individuals



Integrative
Genomics
Viewer



Exclude sequencing and alignment
errors: evaluate sequencing
parameters (GQ, DP, AF) and check
in **IGV**

Variant interpretation

Consensus of pathogenicity
predictors (>5/12)
CADD score (>15)
Conservation score GERP⁺⁺ >3



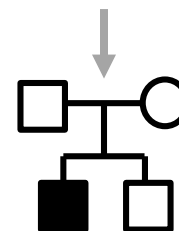
In silico protein
structural - functional
analysis of the
mutation effects



Splicing prediction
Human Splicing Finder
(HSF)

Variant classification

Sanger Validation



Segregation analysis



Transcript analysis
X-chr inactivation

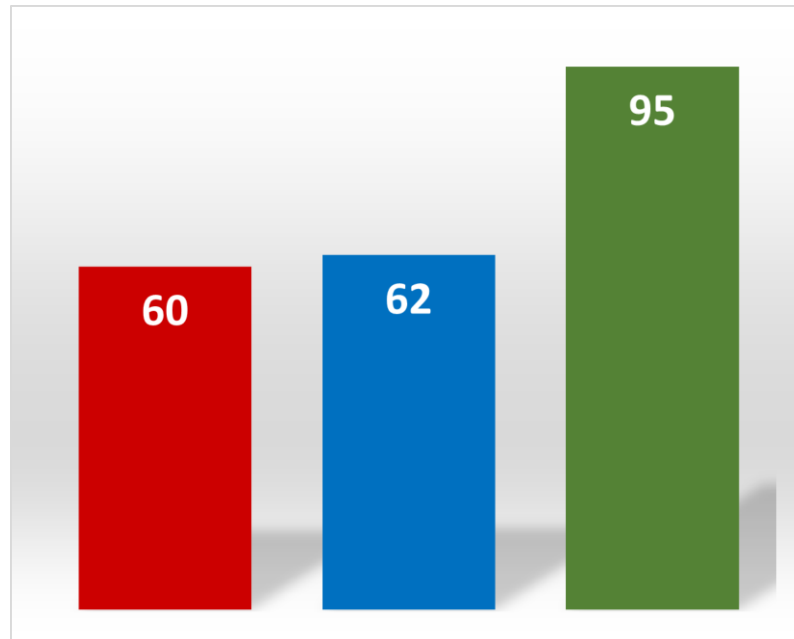


Phenotype consistency



InterVar: Classify System

ID-panel dataset: distribution of DC, LP and CF



■ Disease Causing
■ Likely Pathogenic
■ Contributing Factor

Putative or LP variants need further investigation to be confirmed

Contributing factors are variants in genes associated with ASD

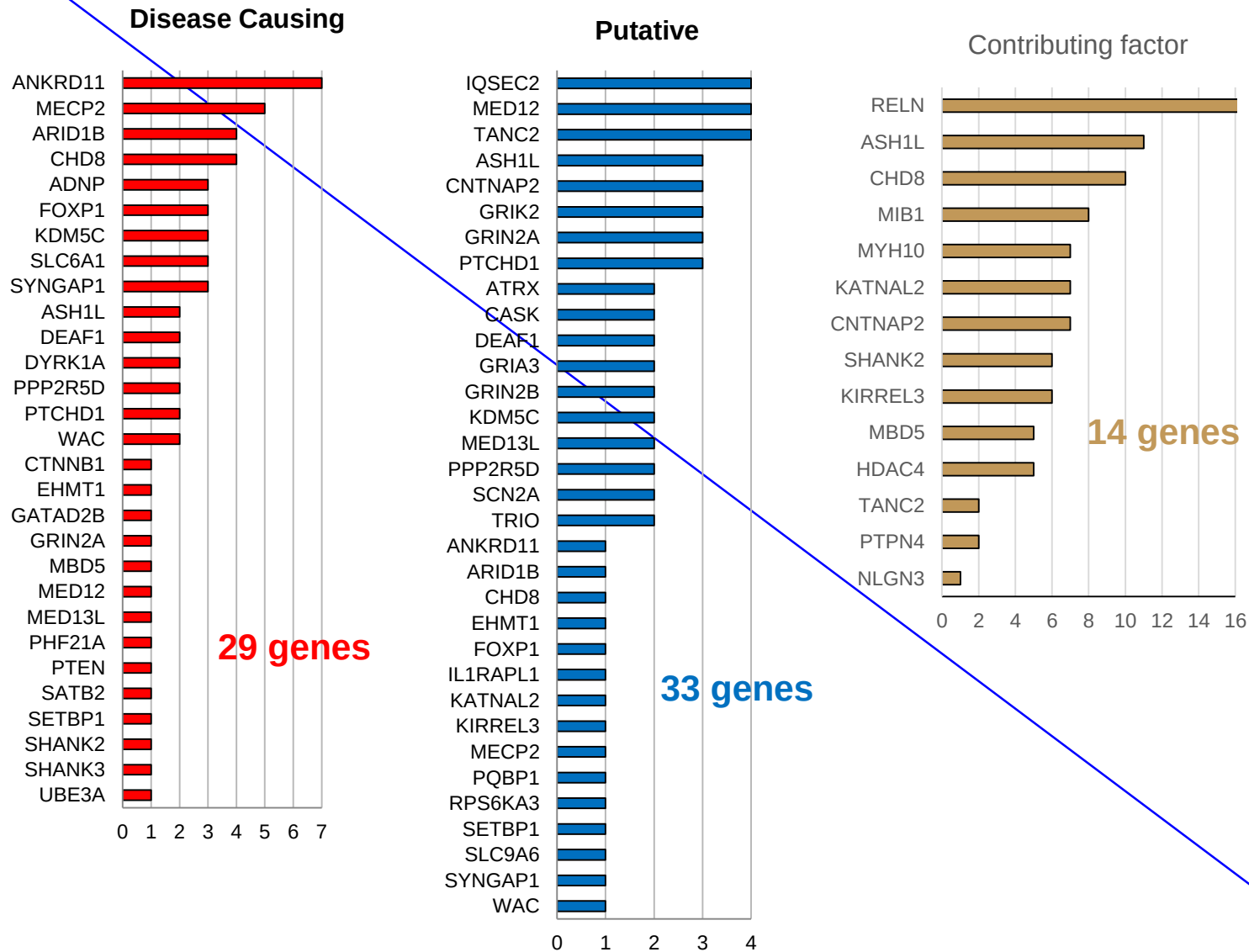
A **confident diagnosis** has been reached in **60 cases (14,5%)**

Some patients carry more than one variant

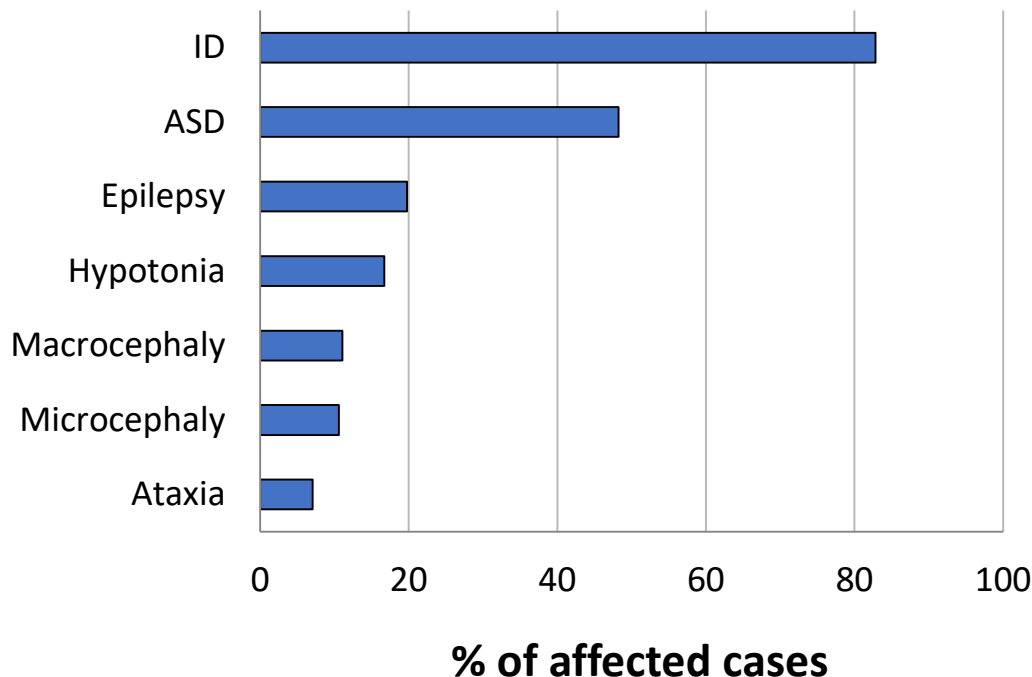
235 (57%) patients with **no** selected variants

Embargoed

ID-panel dataset: distribution of DC, LP and CF



ID-Panel dataset: Clinical features of 415 patients



The clinical features of patients with unspecified NDDs collected using a **personalized medical report**



The median age is 9 years

A patient can be associated with more than one phenotype

N° of selected variants for each phenotype

Phenotype	N° of total cases	% Patients with ≥ 1 SNV	% Disease Causing SNVs	% Putative SNVs	% Contributing Factor
ID	352	46	16	16	22
ASD	205	42	11	14	25
Epilepsy	84	42	18	13	19
Microcephaly	45	51	20	18	18
Macrocephaly	50	42	12	14	16
Hypotonia	72	39	15	14	10
Ataxia	33	39	15	12	12

- ✓ Patients with **at least one selected variant** for each phenotype
- ✓ Each **variant is specific** for each patient

ID panel Dataset: CAGI6

Patient code	ID	ASD	Epilepsy	Microcephaly	Macrocephaly	Hypotonia	Ataxia	Variants
UniPD_0010	1	0	0	0	0	NA	NA	chr16:9934644:C:T
UniPD_0014	1	1	1	NA	NA	1	NA	chr11:45975137:CTGTT:C
UniPD_0047	1	1	0	0	0	0	0	chr3:11067528:G:A
UniPD_0056	1	1	0	NA	NA	NA	NA	chr6:42975009:G:A

- ✓ Information for **7 different phenotypes** to predict for each patients (47% have all phenotype information) (0 = no disease, 1 = disease, NA= not available)
- ✓ Selected **causal variant(s)** from the ID gene panel dataset
- ✓ **Training dataset** (data from CAGI5 ID-panel challenge). Phenotype and variants from 150 patients (Aspromonte et al., 2019)

Prediction template: CAGI6

Patient code	ID-P	ID-SD	ID-V	ID-C	ASD-P	ASD-SD	ASD-V	ASD-C	Epilepsy-P	Epilepsy-SD	Epilepsy-V	Epilepsy-C
UniPD_0001	0.8496	0	3:71021783:T:TA	*	0.643	0	3:71021783:T:TA	*	0.5055	0	3:71021783:T:TA	*
UniPD_0002	0.9053	0	5:14488090:C:G:C	*	0.6002	0	11:694960:CG:C	*	0.5063	0	5:14488090:CG:C	*
UniPD_0003	0.8955	0	2:240098086:T:TC	*	0.664	0	2:240098085:CT:T:C	*	0.4069	0	2:240098086:T:TC	*
UniPD_0005	0.962	0	2:240098086:T:TC	*	0.5942	0	2:240098086:T:TC	*	0.5157	0	2:240098086:T:TC	*

- ✓ **Phenotype Probability score:** [0, 1] range: 0 = no disease, 1 = disease.
- ✓ Binary classification that can be treated independently for each phenotype
- ✓ **Standard deviation (SD),** confidence in the [0,1] range
- ✓ **Variant** associated with the phenotype

Intellectual Disability (ID) panel Challenge

Our goals:

- Assess the ability of computational methods to predict comorbid phenotypes from gene panel data
- Assess the accuracy of computational methods in the prediction of causal variants from a set of real genetic data
- Contribute to improve data analysis in clinical practice
- Possible New genetic findings

Predictors methods

- 8 Teams
- 30 Models

Team	Training	Variants Annotation tools	Filtering by low quality	Filtering by frequency	Filtering by variants impact	Inheritance model	Gene-disease association
1	YES	EVIDENCE 1 VEP 2	?	SNVs >5%	YES	YES	OMIM, ClinVar, Uniprot
2	?	?	?	?	YES	?	?
3	?	VEP 2 REVEL	?	YES	YES	?	ClinVar Phenolyzer REVEL
4	?	VEP SNPs&GO	?	YES <1%	YES	?	?
5	YES	VEP	YES	?	YES: CADD, ESM	?	ClinVar CADD
6	YES	?	?	YES: gnomAD, training dataset	EAM	YES	ClinVar, HPO, DisGenet, Genecard, Pubmed
7	YES	ANNOVAR	?	YES: include only <= 1%	LINSIGHT, MutPred2, REVEL	?	gene_type mutation
8	YES	In-house tool	YES	YES	?	YES	Literature

Phenotype Assessment: metrics

- ✓ ROC curves and **AUCs values**
- ✓ **Bootstrapping** to measure accuracy
- ✓ At each bootstrap, cycle samples are weighted to balance classes
- ✓ Selection of the best model per team, based on **highest AUC values**
- ✓ Measured in **different dataset** (cases with at least one selected variants)
- ✓ **Z-score** for final ranking, to evaluate the ability to predict the **combination of phenotypes**

Assessors



Prof. Silvio Tosatto



Prof. Damiano
Piovesan

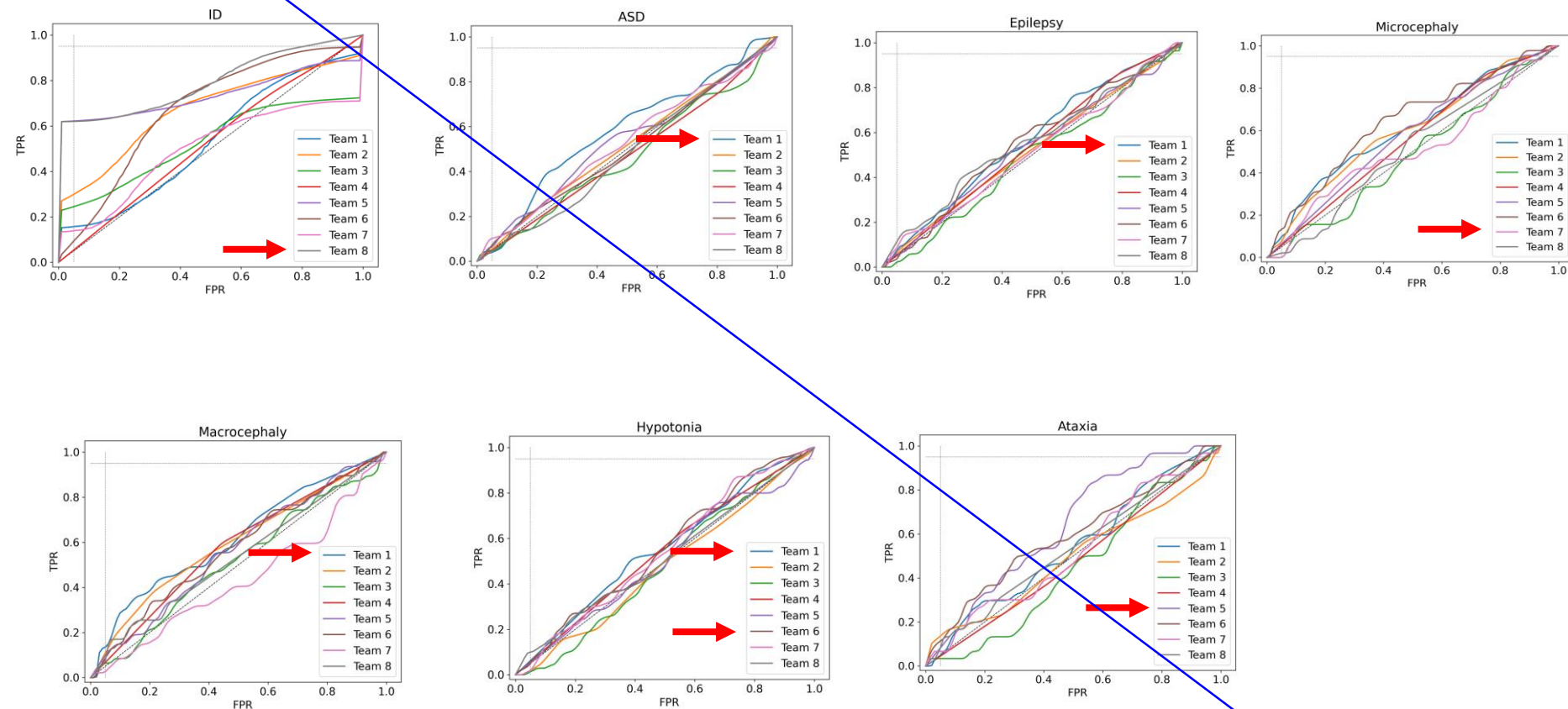


Alessio del Conte,
PhD student

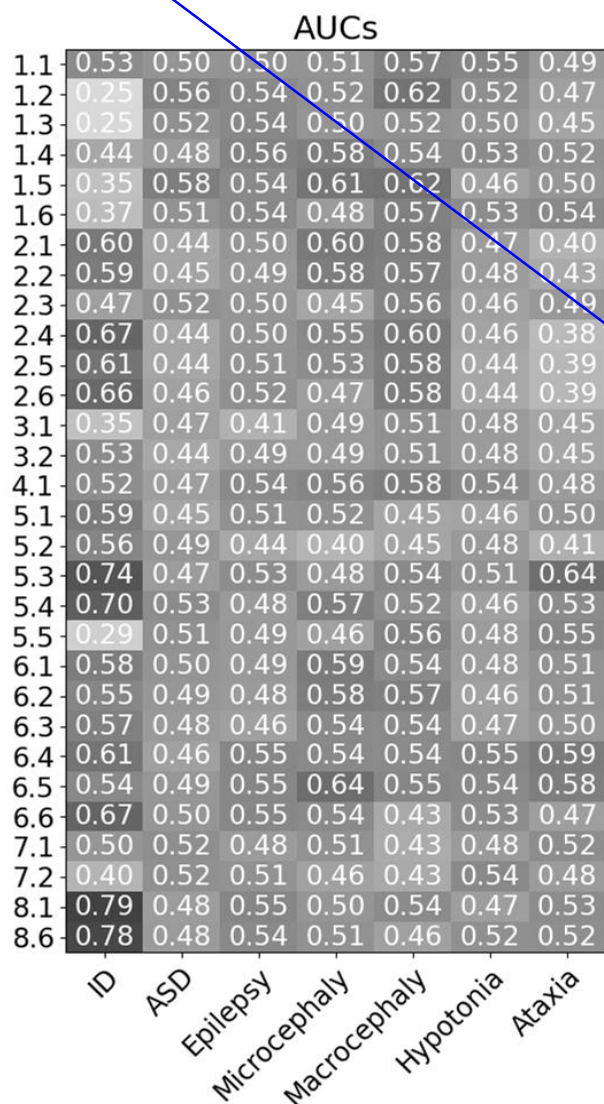


Maria Cristina
Aspromonte, PhD

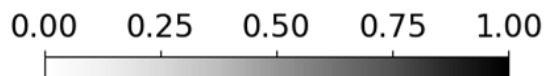
Phenotype Assessment: ROC curves



Phenotype Assessment: AUC values



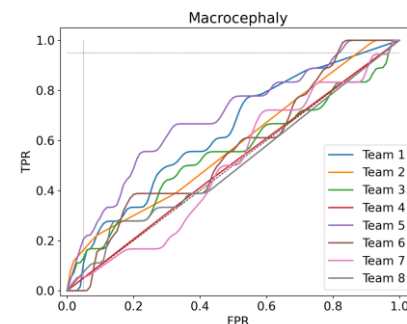
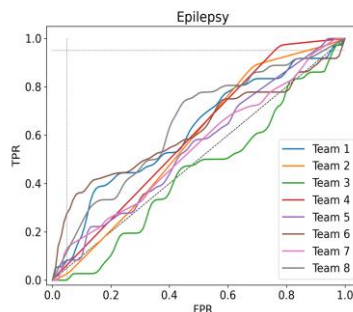
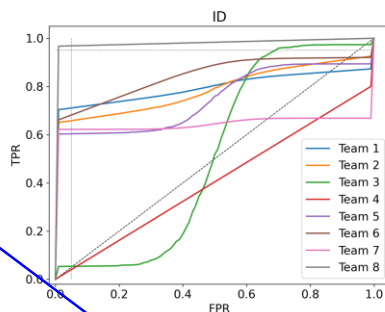
Phenotype	Team model	AUC value
ID	8.1	0.79
ASD	1.5	0.58
Epilepsy	1.4	0.56
Microcephaly	6.5	0.64
Macrocephaly	1.2, 1.5	0.62
Hypotonia	1.1, 6.4	0.62
Ataxia	5.3	0.64



Phenotype Assessment: patients with at least one selected variant

AUCs

1.1	0.79	0.52	0.49	0.56	0.52	0.60	0.50
1.2	0.33	0.55	0.55	0.63	0.63	0.57	0.46
1.3	0.33	0.54	0.55	0.61	0.53	0.55	0.52
1.4	0.48	0.55	0.54	0.61	0.46	0.59	0.39
1.5	0.44	0.57	0.57	0.68	0.63	0.51	0.54
1.6	0.48	0.53	0.62	0.57	0.59	0.51	0.63
2.1	0.67	0.52	0.49	0.51	0.50	0.46	0.29
2.2	0.68	0.51	0.49	0.51	0.48	0.47	0.28
2.3	0.54	0.49	0.52	0.47	0.53	0.45	0.37
2.4	0.79	0.49	0.51	0.45	0.53	0.46	0.35
2.5	0.72	0.48	0.50	0.42	0.51	0.44	0.32
2.6	0.78	0.45	0.58	0.39	0.58	0.49	0.48
3.1	0.51	0.44	0.43	0.42	0.55	0.46	0.40
3.2	0.46	0.47	0.45	0.42	0.55	0.46	0.40
4.1	0.41	0.50	0.60	0.61	0.51	0.55	0.55
5.1	0.67	0.49	0.55	0.45	0.49	0.47	0.49
5.2	0.59	0.48	0.48	0.33	0.48	0.48	0.41
5.3	0.75	0.51	0.54	0.49	0.60	0.47	0.65
5.4	0.55	0.55	0.47	0.65	0.55	0.43	0.46
5.5	0.21	0.50	0.47	0.46	0.69	0.51	0.66
6.1	0.53	0.55	0.55	0.62	0.51	0.48	0.51
6.2	0.51	0.46	0.53	0.58	0.52	0.49	0.53
6.3	0.65	0.51	0.52	0.58	0.56	0.45	0.54
6.4	0.78	0.51	0.52	0.51	0.50	0.52	0.48
6.5	0.57	0.52	0.60	0.68	0.49	0.50	0.48
6.6	0.84	0.48	0.62	0.62	0.47	0.51	0.44
7.1	0.64	0.53	0.46	0.50	0.50	0.44	0.58
7.2	0.40	0.54	0.55	0.38	0.49	0.48	0.55
8.1	0.98	0.45	0.63	0.61	0.51	0.51	0.58
8.6	0.96	0.47	0.64	0.59	0.49	0.52	0.58
ID ASD Epilepsy Microcephaly Macrocephaly Hypotonia Ataxia							



	All cases		Cases with at least one selected variant	
Phenotype	Best Team model	AUC value	Best Team model	AUC value
ID	8.1	0.79	8.1	0.98
ASD	1.5	0.58	1.5	0.57
Epilepsy	1.4	0.56	8.6	0.64
Microcephaly	6.5	0.64	1.5, 6.5	0.68
Macrocephaly	1.2, 1.5	0.62	5.5	0.69
Hypotonia	1.1, 6.4	0.62	1.1	0.60
Ataxia	5.3	0.64	5.5	0.66

Phenotype Assessment: combined

Z-score was used on AUCs values to compute ranking

Team.Model	ID	ASD	Epilepsy	Microcephaly	Macrocephaly	Hypotonia	Ataxia	Mean	Model Rank	Team Rank
1.5	-1,18	2,61	0,75	1,50	1,55	-0,97	0,09	0,62	3	2
2.2	0,40	-1,17	-0,52	1,08	0,58	-0,38	-0,90	-0,13	16	6
3.2	-0,06	-1,38	-0,72	-0,73	-0,38	-0,29	-0,63	-0,60	28	8
4.1	-0,08	-0,44	0,83	0,56	0,77	1,28	-0,11	0,40	7	4
5.3	1,44	-0,40	0,52	-0,87	0,20	0,42	2,50	0,55	4	3
6.5	0,05	0,16	1,06	2,17	0,33	1,40	1,56	0,96	1	1
7.2	-0,93	0,85	0,04	-1,22	-1,89	1,47	-0,24	-0,27	22	7
8.1	1,77	-0,40	1,20	-0,45	0,00	-0,70	0,74	0,31	9	5

Variants Assessment: metrics

Predictor assigned zero or more variants to each patient

If a variant is assigned multiple times to a patient it was counted as one

- Recall (sensitivity)
- Accuracy
- Selection of the best model per team, based on recall and accuracy
- Assessment on different sets of data (class, type, Clinvar)
- Comparison with CADD score (>25 cut-off used)
- Comparison with consensus predictions from ANNOVAR

Variants assessment

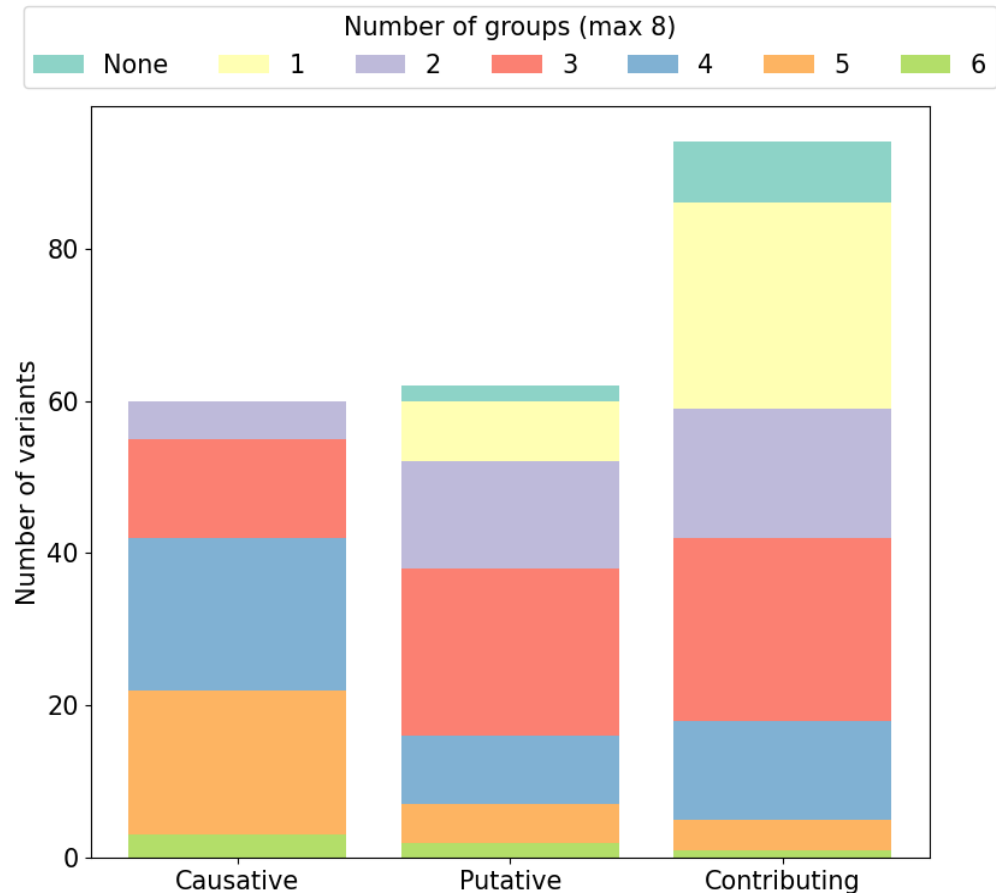
Model	Correctly pred. variants	Total pred. DC variants	Total pred. variants	Correctly pred variants/Exp variants (Recall)	Correctly pred variants/tot pred variants (Accuracy)	Correctly pred variants (DC)/Exp variants (DC)
1.1	20	14	628	0.1	0.03	0.2
2.5	3	2	182	0.02	0.02	0.03
3.1	115	36	233	0.59	0.49	0.6
4.1	61	38	251	0.31	0.24	0.6
5.1	19	7	182	0.1	0.1	0.1
6.3	91	51	129	0.46	0.71	0.9
7.1	114	36	294	0.58	0.39	0.6
8.6	167	54	299	0.85	0.56	0.9

For each model a count of the variants is done, assessing the accuracy and recall

Variants Assessment

Consensus of groups over variants,
divided **by class**

- ✓ Causative variants are the ones with most consensus
- ✓ 60% Putative and Contributing Factor variants were predicted by at least two groups

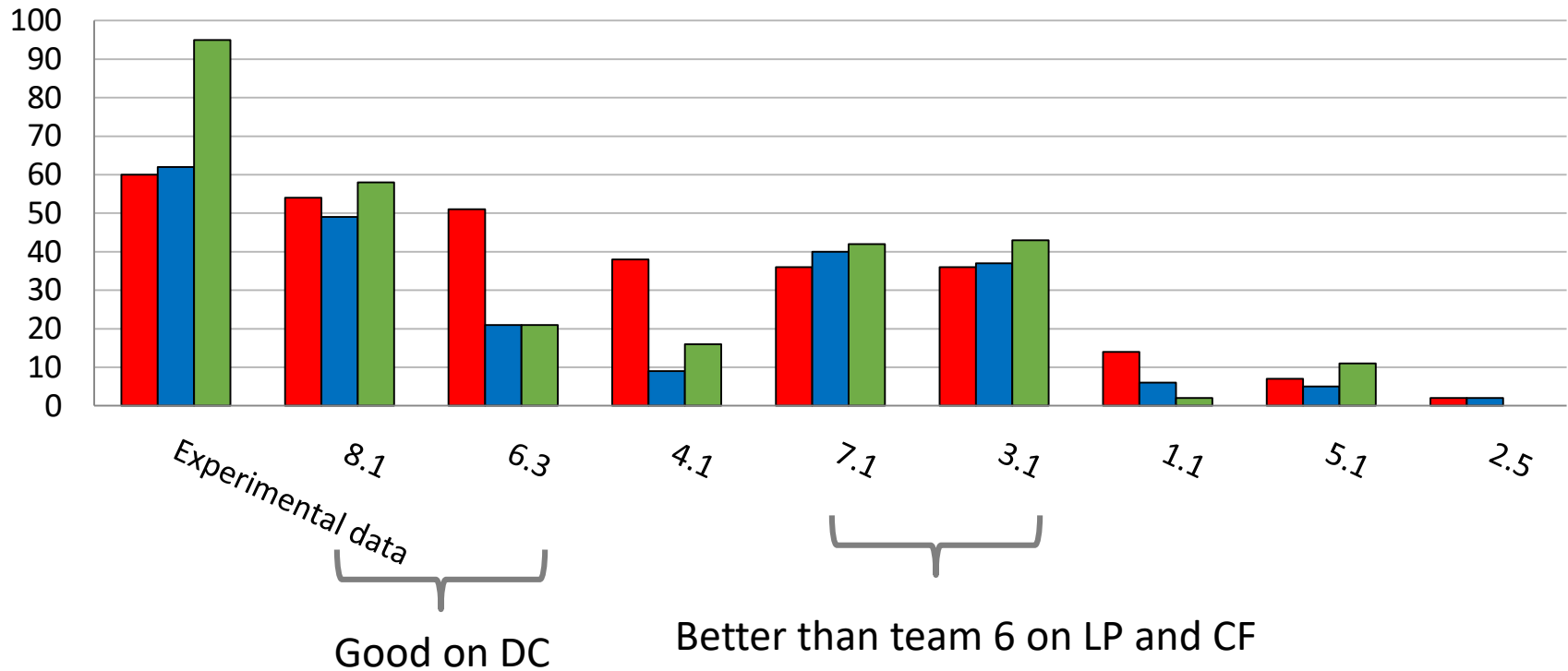


Variants Assessment

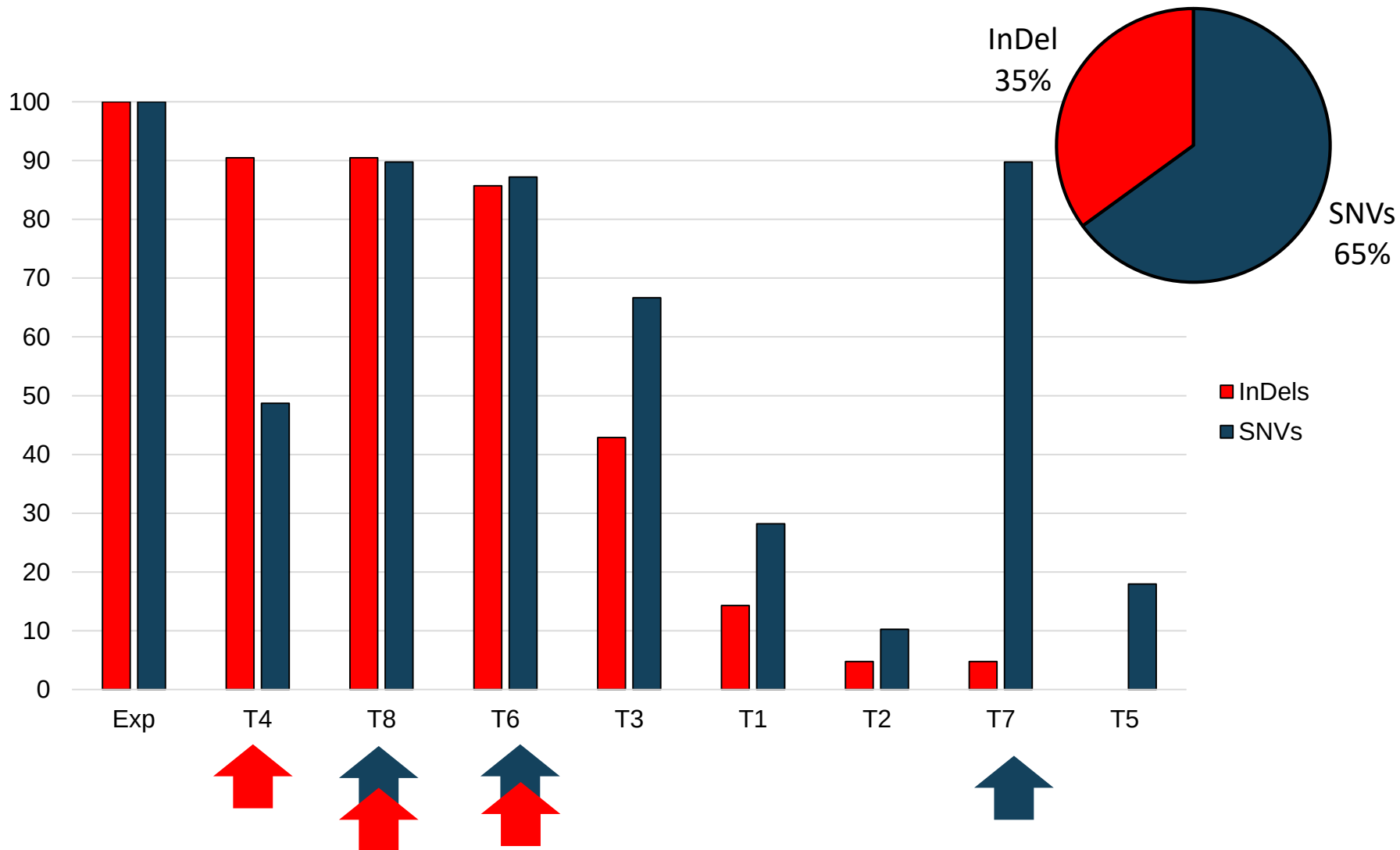
on diverse classes of selected variants

N° of Correctly Predicted
Variants

■ Disease Causing ■ Likely Pathogenic ■ Contributing Factor

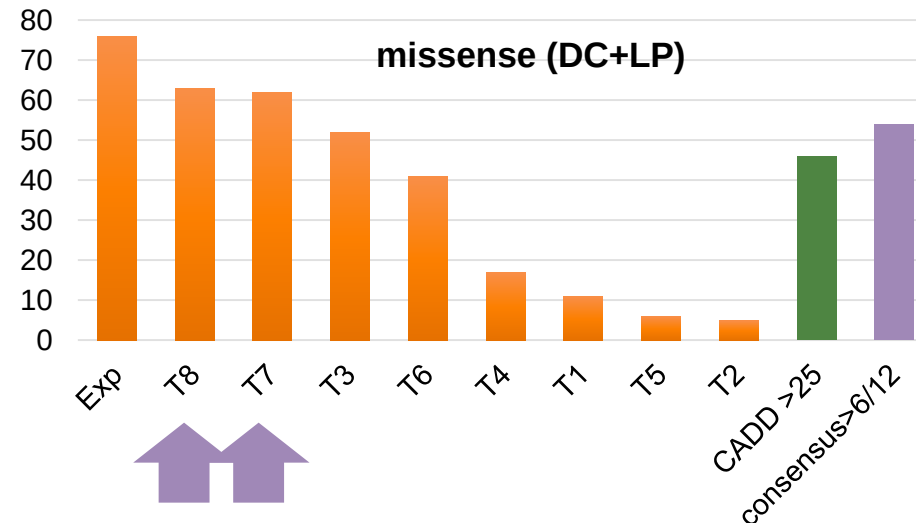
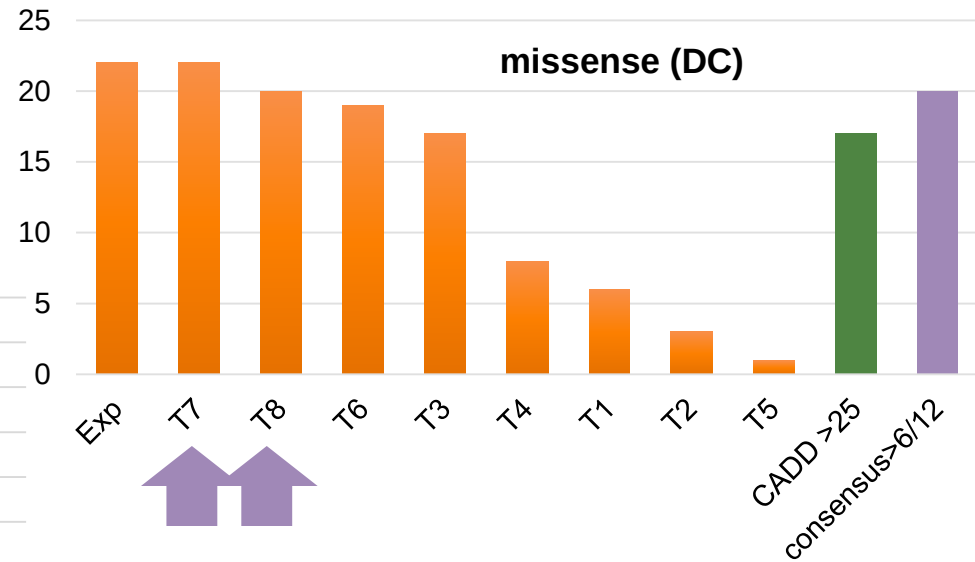
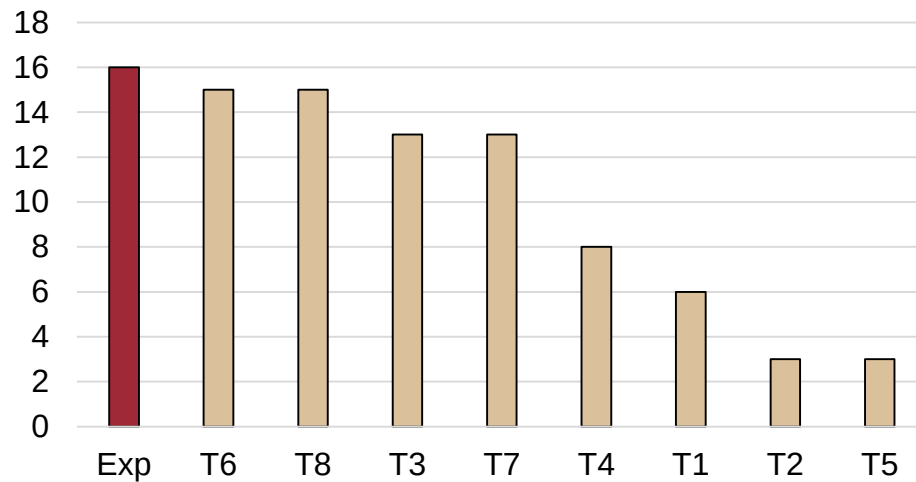


Performance on different DC variant types



Variants Assessment: Clinvar and missense impact

ClinVar=Pathogenic, Likely Pathogenic



Some DC variants Hard to predict

all DC variants have been predicted by at least 2 teams

Patient code	Func.refGene	Chr position	variant	Predicted by	
Deep Splicing variants					
UniPD_0383	Intronic	chr12:116420425:T:C	MED13L:NM_015335.4:c.4956-17A>G	T1, T2, T5, T8	Alters trasncript
Missense Variants with CADD<25 and D <6/12 tools					
UniPD_0169	exonic	chrX:23411707:A	PTCHD1:NM_173495:exon3:c.G2072A:p.R691Q	T3, T7	23.3, D 5/12
UniPD_0286	exonic	chr1:153789043:A:C	GATAD2B:NM_020699:exon7:c.T922G:p.C308G	T1, T3, T4, T6, T7, T 8	12.5, D 3/12
UniPD_0337	exonic	chr1:155449482:T:C	ASH1L:NM_018489:exon3:c.A3179G:p.N1060S	T6, T7, T8	23.1
Variants with BAD sequenicng parameters					
UniPD_0010	exonic	chr16:9934644:C:T	GRIN2A:NM_000833:exon8:c.G1511A:p.R504Q (mosaic)	T1, T3, T7, T8,	GQ: 31, DP: 310, AF: 0.16
UniPD_0418	exonic	chr11:70331632:GCACTGCCCGGC GGGGG:G	SHANK2:NM_133266:exon10:c.2986_3001del:p.P996fs	T4, T8	GQ: 7, DP: 11, AF: 0.7

Variant assessment: Putative variants

60% of the putative variants were supported by at least two Teams

NOT PREDICTED

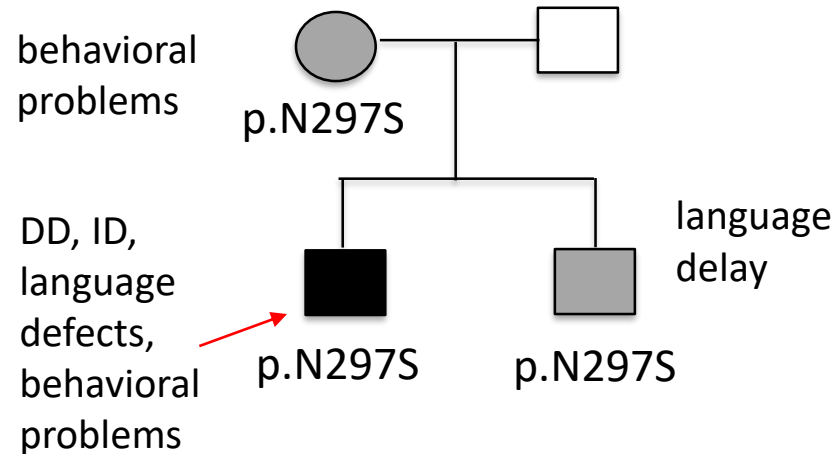
UniPD_0273: CNTNAP2:NM_014141:exon16:c.T2517C:p.N839N

Paternal, in trans with the maternal CNTNAP2:NM_014141:exon4:c.C511T:p.R171C

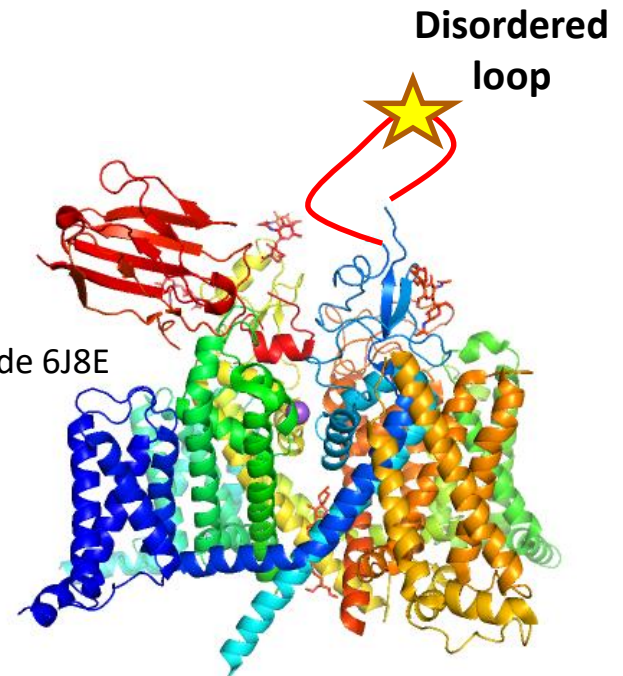
UniPD 0386: SCN2A:NM_021007:exon7:c.A890G:p.N297S

Not in Gnomad, Clinvar VUS, CADD 8.3, damaging by 9/12 tools

Ion transport domain

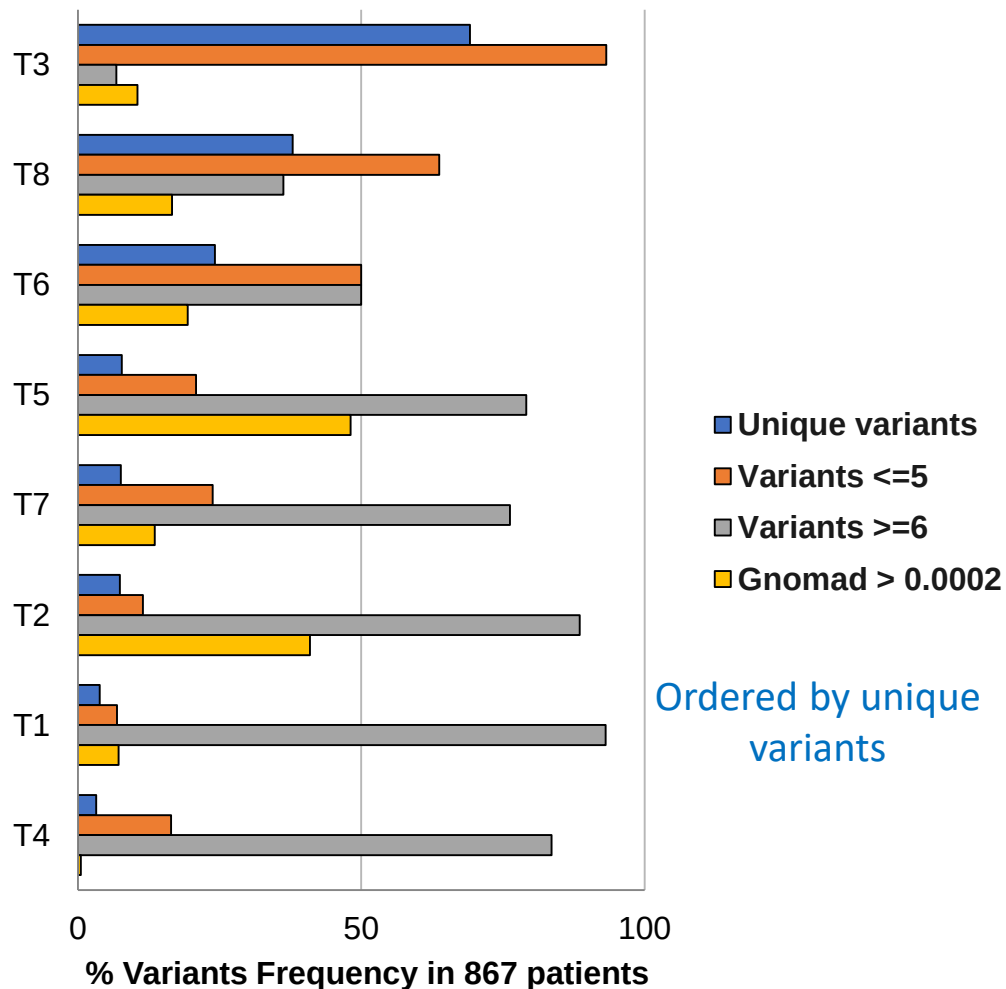


SCN2A
PDB code 6J8E



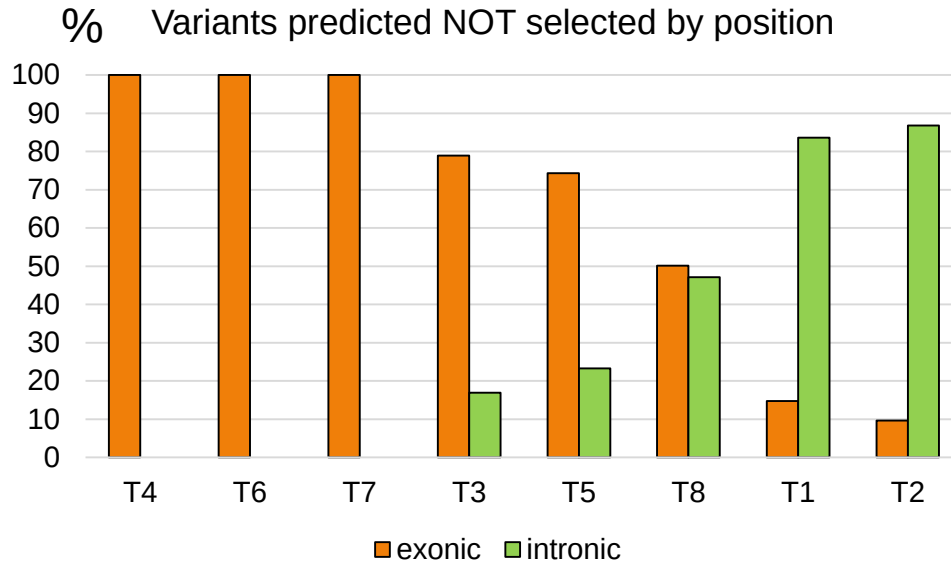
Variants re-evaluation

Variants Predicted NOT Selected



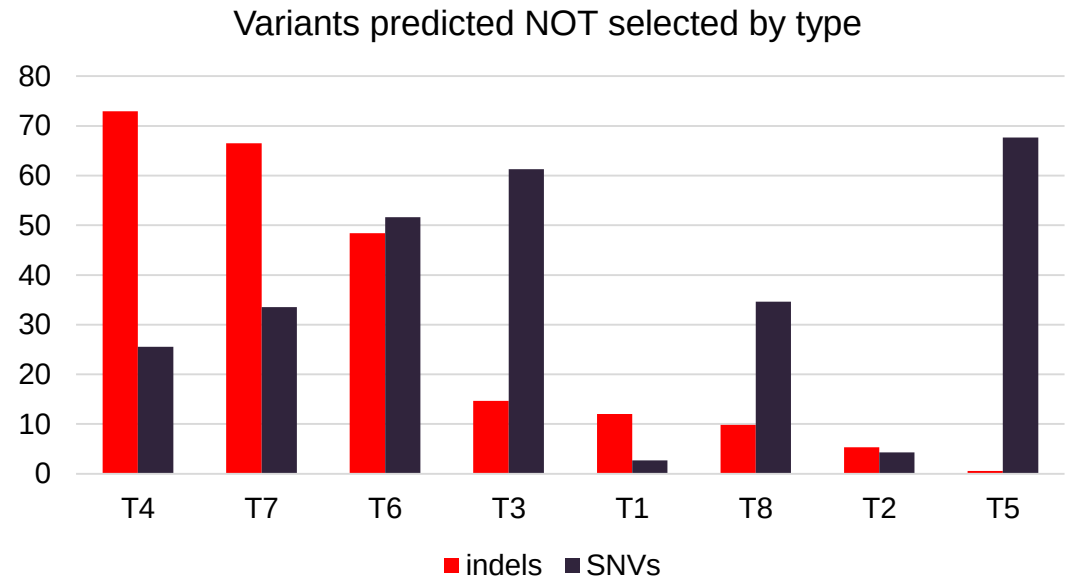
- Groups T3 and T8 reported a high % of unique or rare variants
- Other groups (T1 and T4) selected variants frequently found in our cohort, however this does not correlate with the frequency of these variants in the general population
- This means that these variants could be **technical errors**

Variants re-evaluation



Most of the variants reported by T1 and T2 are intronic variants

Most of the variants reported by T4 and T7 are indels variants, while T5 reported almost exclusively SNVs



Variants re-evaluation

Filtering

8794

1411

824

~400

~80

173

- Variants predicted by the CAGI teams (Not selected by us)
- Re-evaluated variants found <5/867
- Unique variants (1/867) - 526 exonic,
• 262 intronic
- Check raw data with IGV: **50 %** of errors
- Sanger validation: confirmed errors
- Segregation analysis: inherited from healthy parents

New finding: AP1S2

UniPD_0215 (M, 17y)

chrX:15864074:C:T

AP1S2:NM_003916:exon3:c.G240A:p.L80L

Not in GnomAD

2/867 (one is a female; usually AP1S2

female carriers are healthy)

Predicted by Team 8 and 1

Features of Pettigrew syndrome ([304340](#))

- Intellectual disability, severe
- Seizures
- Ataxic gait
- Self-injurious behavior
- facial dysmorphism: large mouth
- MRI: enlarged ventricles

Impact Prediction

Format: text Download Copy

Type	Interpretation		
Alteration of auxiliary sequences	Significant alteration of ESE / ESS motifs ratio (-6)		
Algorithm/Matix	position	sequence	
ESS_hnRNPA1 (New ESS Site)	chrX:15845952	TAGAAA	
RESCUE ESE (ESE Site Broken)	chrX:15845952	TGGAAA	
Sironi_motif1 (New ESS Site)	chrX:15845953	CTAGAAAT	
ESE_ASFB (ESE Site Broken)	chrX:15845955	CCCTGGA	
ESE_SC35 (ESE Site Broken)	chrX:15845958	TTACCTG	
PESS (New ESS Site)	chrX:15845958	TTACCCTA	
New Acceptor splice site	Activation of a cryptic Acceptor site. Potential alteration of splicing		
Algorithm/Matix	position	sequences	variation
HSF Acceptor site (matrix AG)	chrX:15845961	- REF : TAATTACCCTGGAA - ALT : TAATTACCCTAGAA	49.16 > 77.03 => 56.69%

We are waiting for parents DNA samples and proband sample for RNA

Conclusions

Phenotype Assessment

- **Phenotype is hard to predict** with gene panel data
- **Team 8** reached the best AUC value on ID prediction. Other team models (**1, 5, 6, 7**) performed better in specific phenotypes
- **Team 6** best performance with Z-score, followed by **Team 1**

Variant Assessment

- **Team 8** overall winners, followed by **Team 6**
- The recognition of **technical errors** is an important step to avoid noisy
- **New Findings:** A novel **putative synonymous variant in AP1S2 gene**, consistent with the patient phenotype
- Improve automated **splicing prediction** in clinical practice

Acknowledgements

PROVIDERS

Molecular Genetics of Neurodevelopmental Disorders Lab, Dept. Woman and Child Health, University of Padova, Italy

Alessandra Murgia

Maria Cristina Aspromonte

Italian clinical and genetic centers referring patients

FAMILIES



ASSESSORS

BioComputing Group (Dept. Biomedical Sciences) University of Padova, Italy

Silvio Tosatto

Alessio del Conte

Damiano Piovesan



PREDICTORS:

AIBI-CAGI6, BioStat_CUHK, Bologna Biocomputing Group, Hoehndorf, Lichtarge, Mooney-Radivojac, ILHyd

CAGI ORGANIZERS

Steven Brenner, Pedja Radivojac, Tina Bakolitsa

