Frederick
National Laboratory

NIH NATIONAL CANCER INSTITUTE

Statistical approaches to detect pathogenic variants of the *BRCA2* oncogene

Alexander Y. Mitrophanov, PhD

alex.mitrophanov@nih.gov

Advanced Biomedical Computational Sciences (ABCS)
Frederick National Laboratory for Cancer Research

ABCS “Statistics for Lunch,” 8 October 2024

Biology + Statistics

Focus: statistics as an essential means of solving a biological problem

“Disclaimer:” we present a statistician’s perspective!

Biology + Statistics

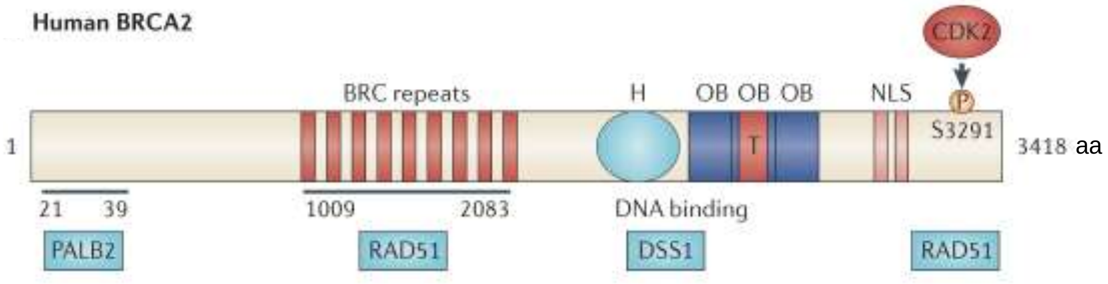
Focus: statistics as an essential means of solving a biological problem

“Disclaimer:” we present a statistician’s perspective!

<u>Full details (including all the biology):</u>

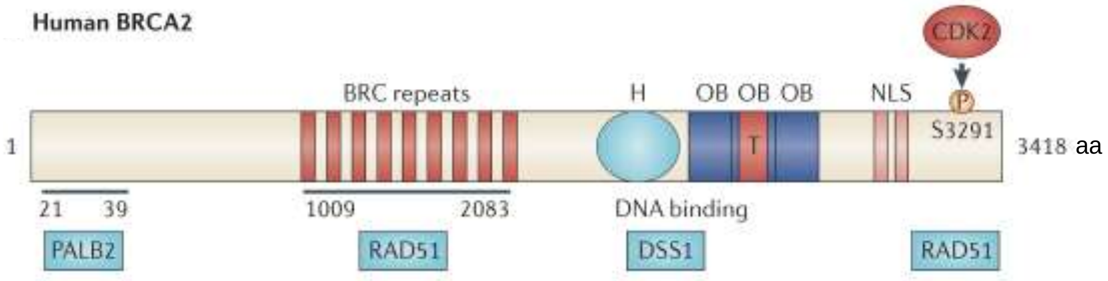
- Biswas, Mitrophanov, …, Sharan (2023) *Cell Rep Methods* **3:** 100628
- Sahu, Sullivan, Mitrophanov, …, Sharan (2023) *PLOS Genet* **19:** e1010940

Pathogenicity of *BRCA2* oncogene variants needs assessment

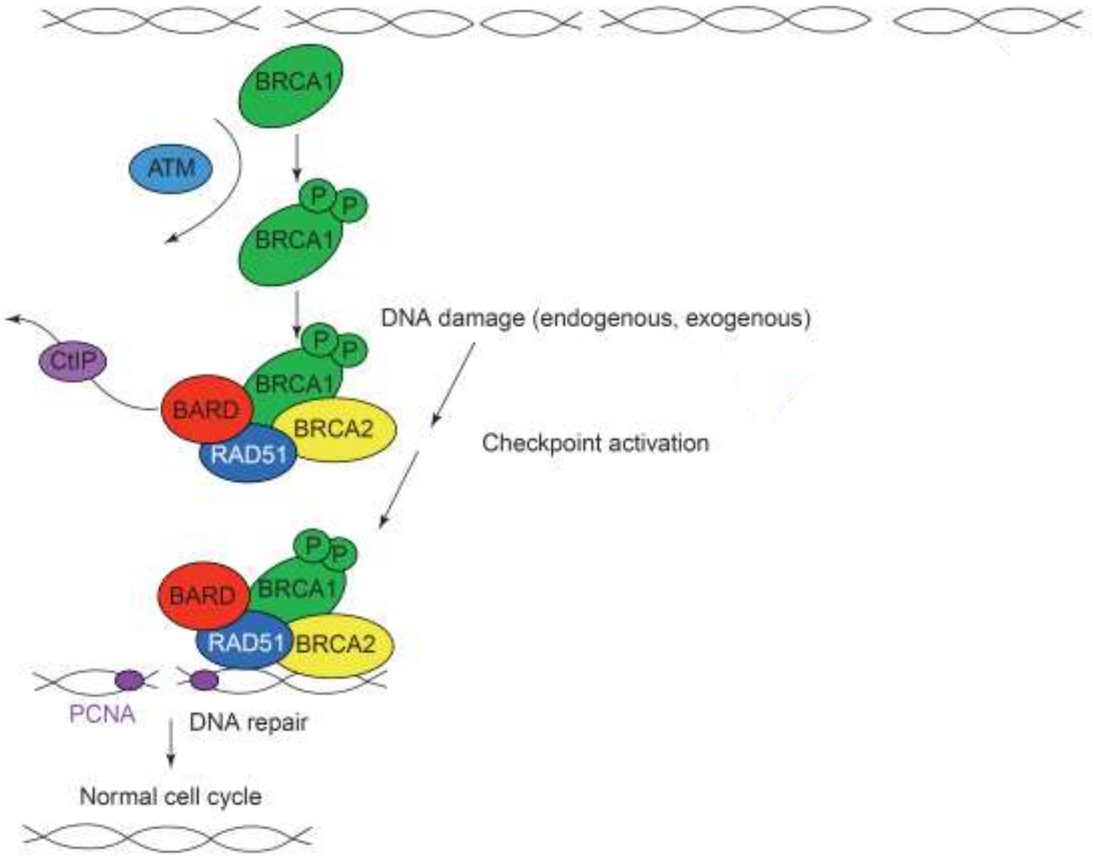


Roy *et al*., *Nat Rev Cancer* 2012

Pathogenicity of *BRCA2* oncogene variants needs assessment

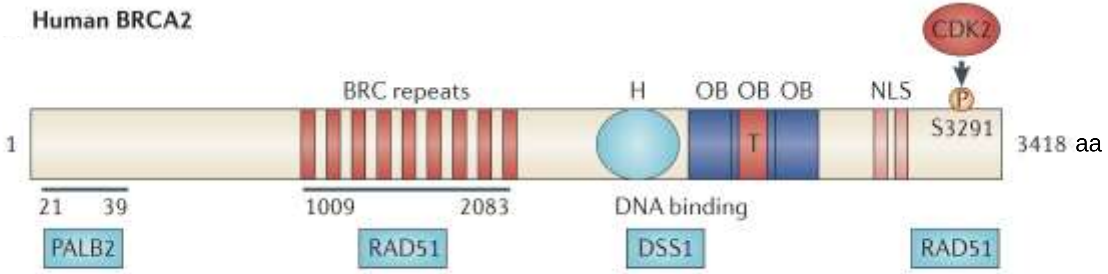


Roy *et al.*, *Nat Rev Cancer* 2012

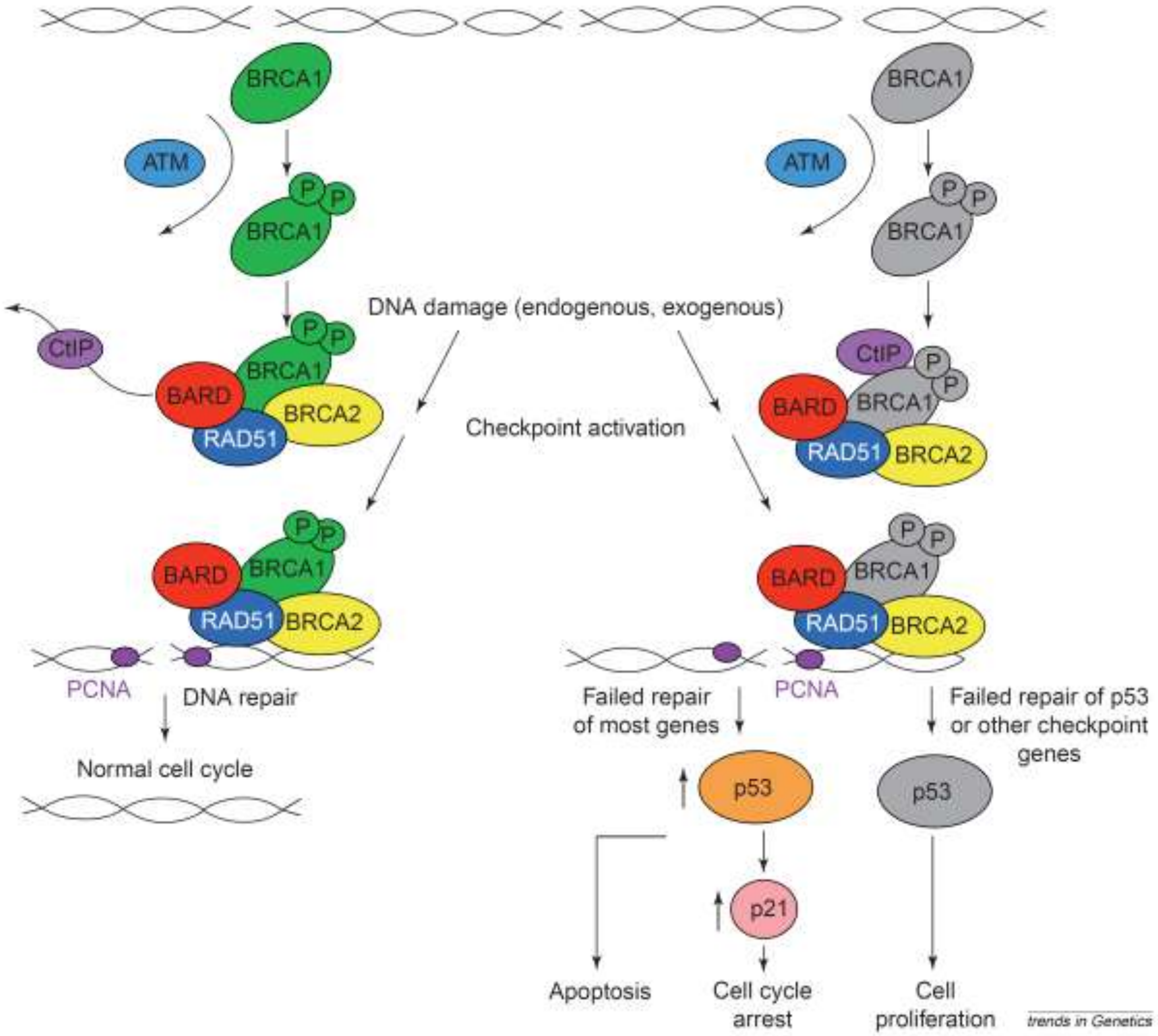


Welcsh *et al.*, *Trends Genet* 2000

Pathogenicity of *BRCA2* oncogene variants needs assessment

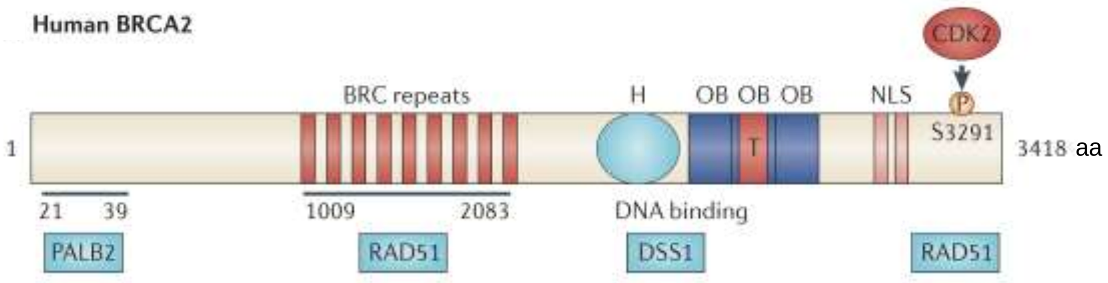


Roy *et al.*, *Nat Rev Cancer* 2012



Welcsh *et al.*, *Trends Genet* 2000

Pathogenicity of *BRCA2* oncogene variants needs assessment



Roy *et al.*, *Nat Rev Cancer* 2012

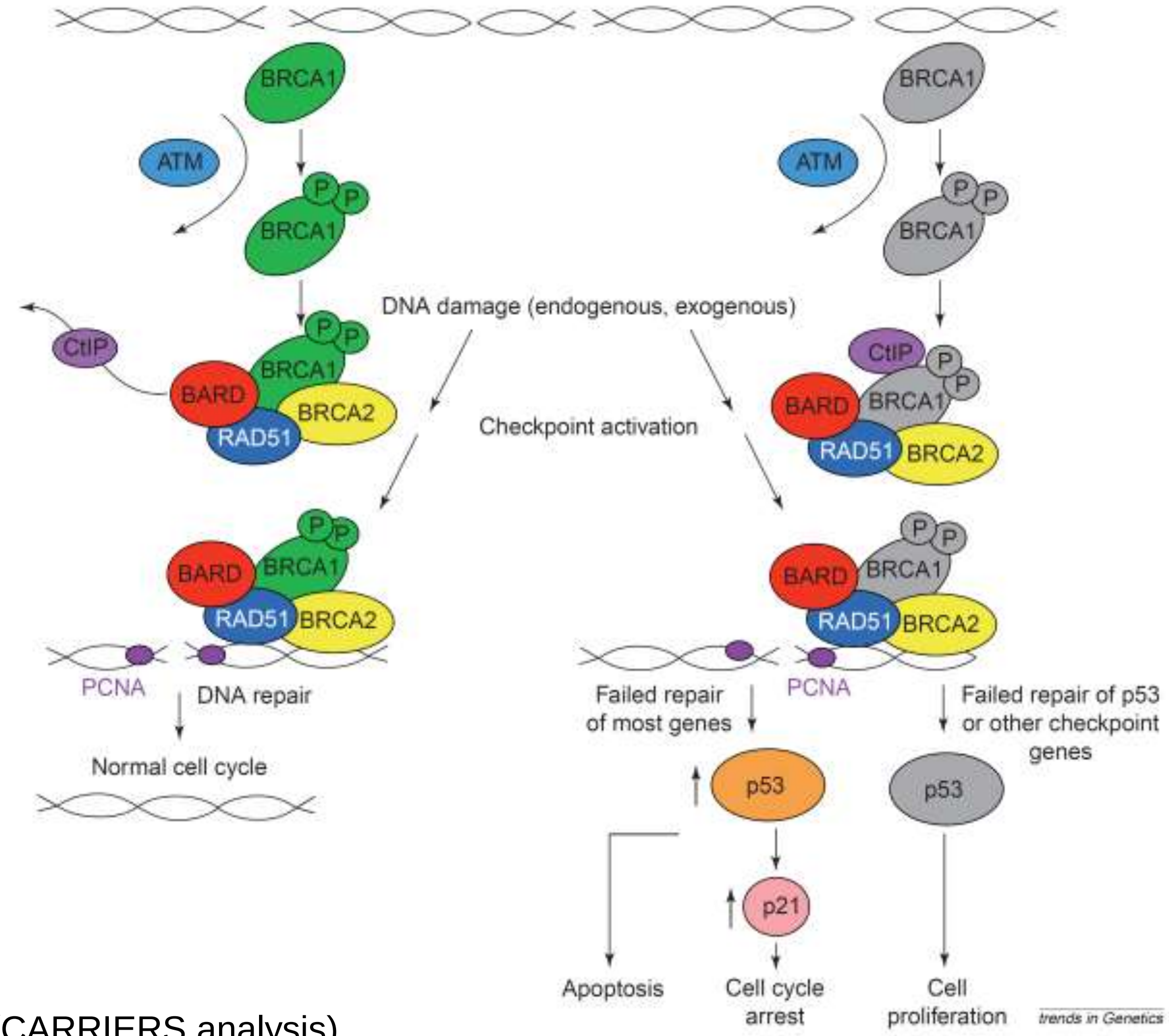
BRCA2 sequence mutations

⇩

Different *BRCA2* variants

⇩

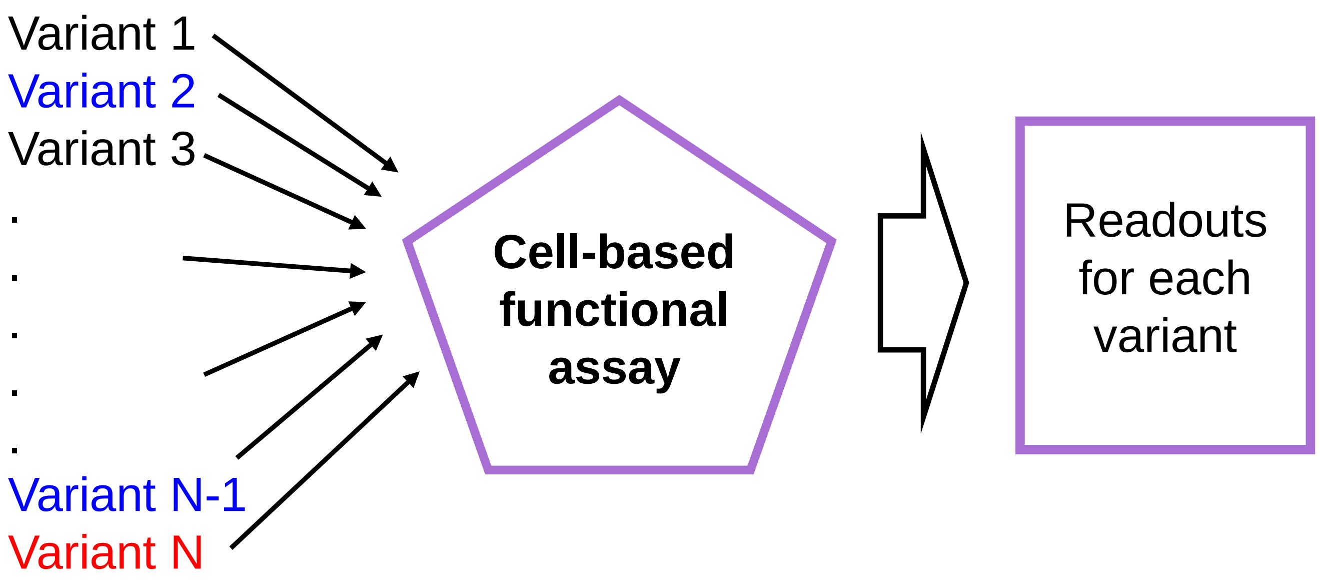
Predisposition to breast and ovarian cancer



~1.3% breast-cancer patients have pathogenic *BRCA2* variants (CARRIERS analysis)
17,000+ *BRCA2* variants in ClinVar database; **3,000+ are of uncertain significance**

Welcsh *et al.*, *Trends Genet* 2000

How do we know which *BRCA2* variants are likely to be pathogenic? By using a functional assay!



Variant 1
Variant 2
Variant 3
.
.
.
.
.
Variant N-1
Variant N

Cell-based functional assay

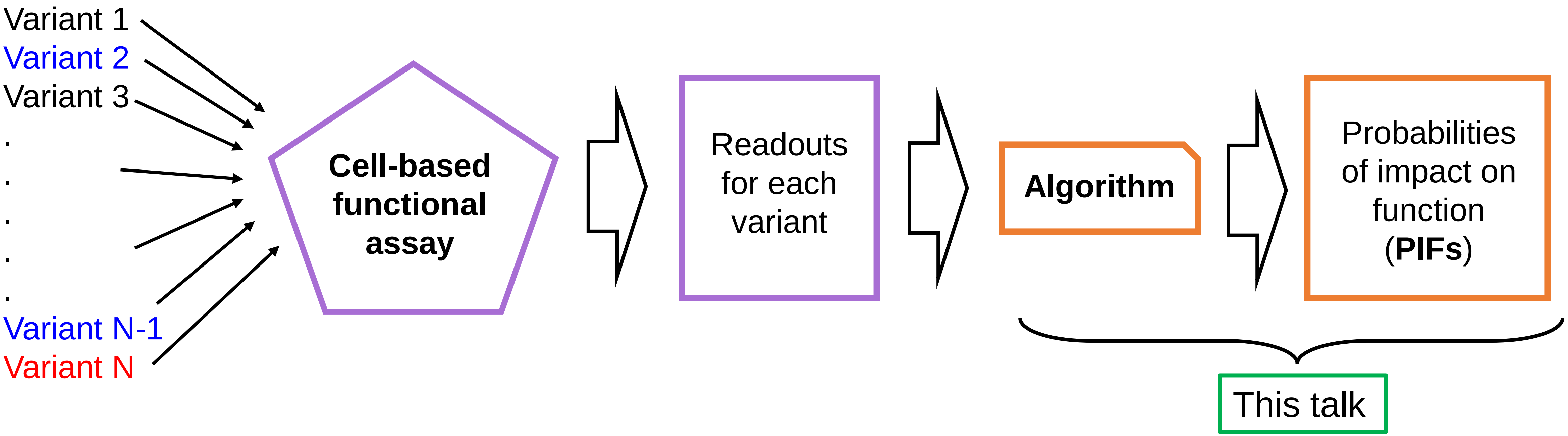
Readouts for each variant

Red: pathogenic variant
Blue: benign variant
Black: VUS (variant of uncertain significance)

American College of Medical Genetics and Genomics:
“A well-established functional assay is *strong* evidence to classify variants”

Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

How do we know which *BRCA2* variants are likely to be pathogenic? By using a functional assay!



Variant 1
Variant 2
Variant 3
.
.
.
.
.
Variant N-1
Variant N
Cell-based functional assay
Readouts for each variant
Algorithm
Probabilities of impact on function (PIFs)
This talk

Red: pathogenic variant

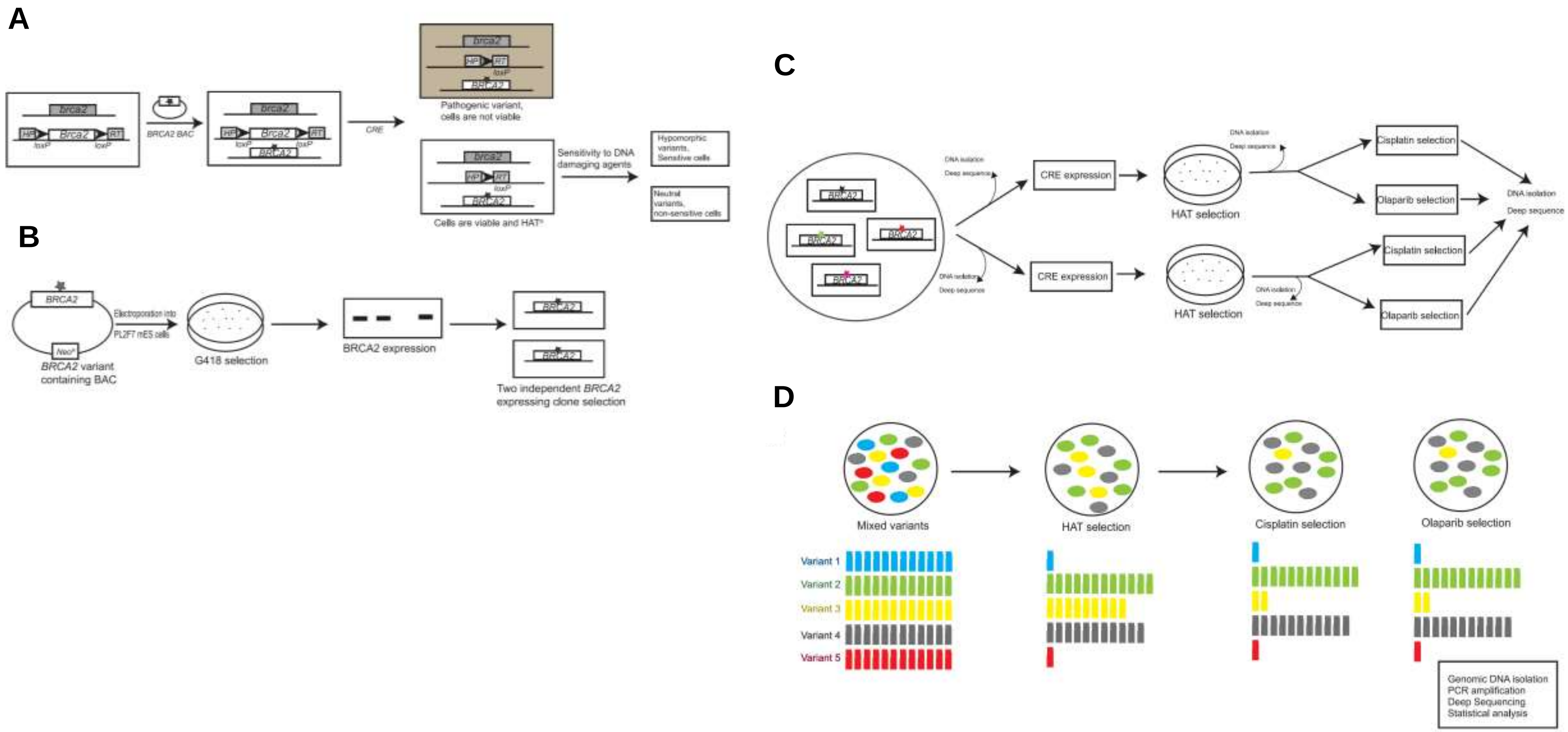
Blue: benign variant

Black: VUS (variant of uncertain significance)

American College of Medical Genetics and Genomics: "A well-established functional assay is *strong* evidence to classify variants"

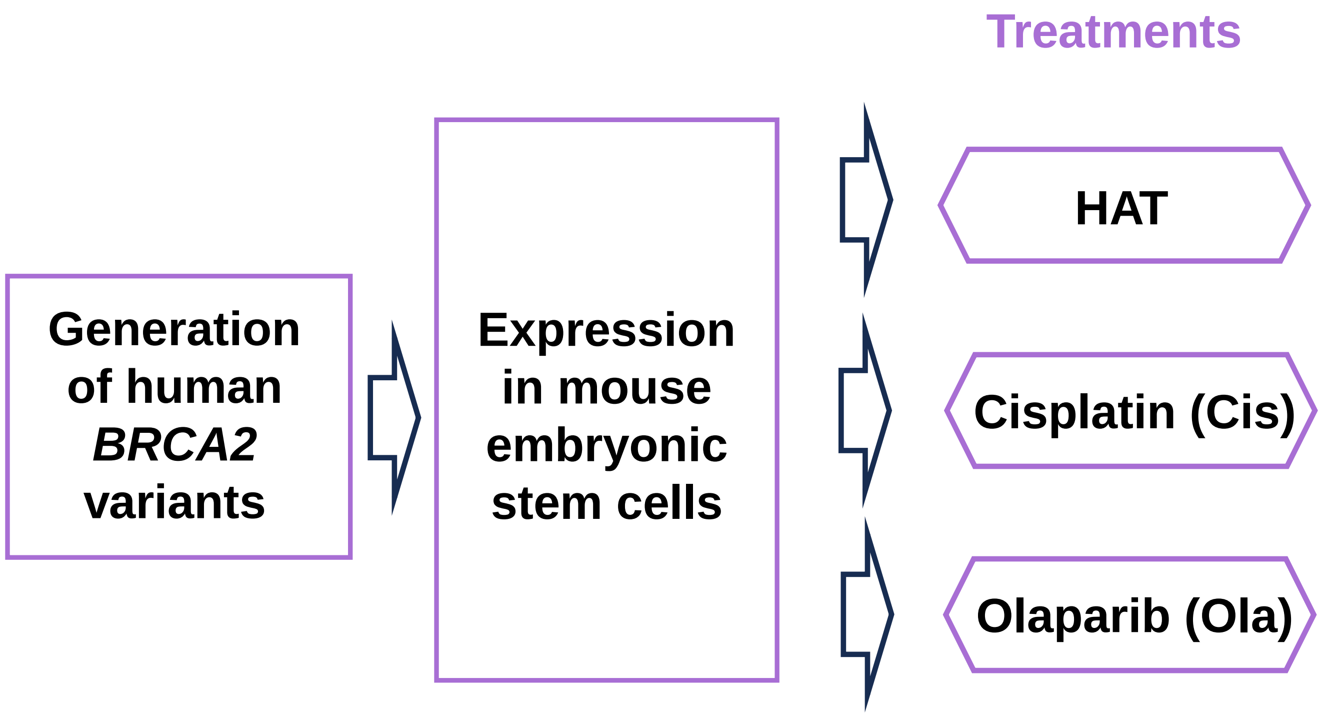
Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Experimental setup for data generation for 223 *BRCA2* variants



Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Experimental setup for data generation for 223 *BRCA2* variants (*streamlined*)



Treatments

Generation of human *BRCA2* variants

Expression in mouse embryonic stem cells

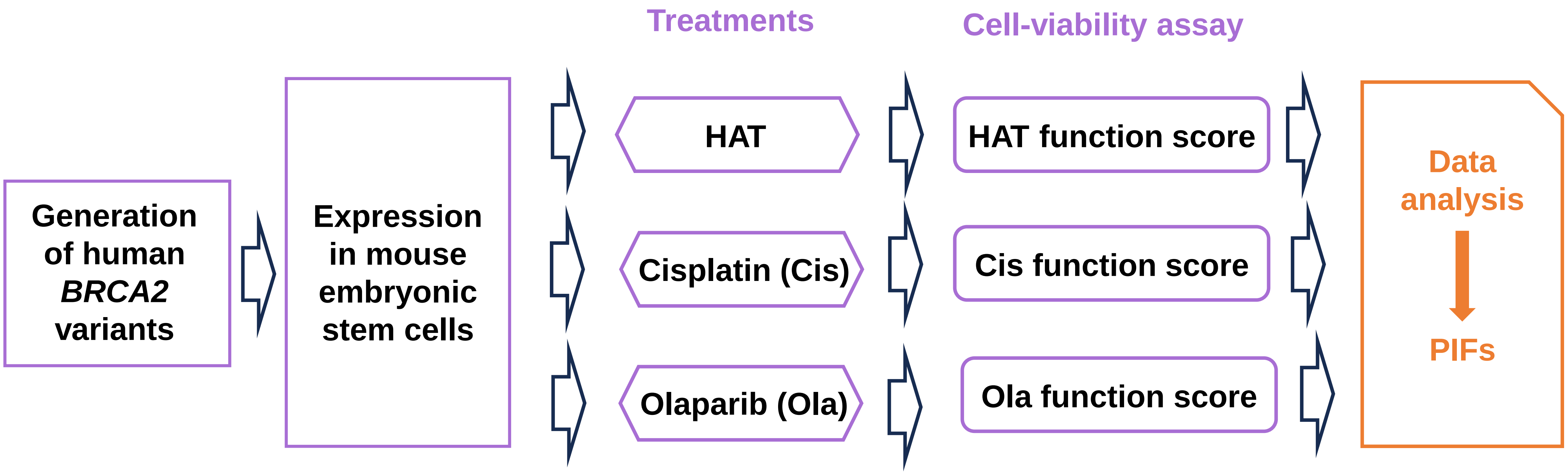
HAT

Cisplatin (Cis)

Olaparib (Ola)

Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

Experimental setup for data generation for 223 *BRCA2* variants (*streamlined*)



Treatments

Cell-viability assay

Generation of human *BRCA2* variants

Expression in mouse embryonic stem cells

HAT

Cisplatin (Cis)

Olaparib (Ola)

HAT function score

Cis function score

Ola function score

Data analysis

PIFs

Function score: frequency of variants in the final pool relative to the initial pool (determined via *next-generation sequencing*)

One data point: the function score for one **variable** (HAT or Cis or Ola) for one *BRCA2* variant

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Requirements for the statistical methodology

Requirements

- Should calculate probabilities of impact on function (PIFs)

- The probabilities should not be “too binary”

- Should use the accepted PIF thresholds for benign and pathogenic (**≤0.05** and **>0.99**, respectively)

- Should use **semi-supervised learning** (expected to outperform supervised-learning approaches; e.g., VarCall software)

Requirements for the statistical methodology

Requirements

- Should calculate probabilities of impact on function (PIFs)

- The probabilities should not be “too binary”

- Should use the accepted PIF thresholds for benign and pathogenic (**≤0.05** and **>0.99**, respectively)

- Should use **semi-supervised learning** (expected to outperform supervised-learning approaches; e.g., VarCall software)

- **Supervised:** use only labeled data in model training (fitting)
- **Semi-supervised:** use all available data in model training (fitting)

Model assessment in statistics

Descriptive and inferential statistics

How well the model captures the statistical *distributions* in the data set; how well it allows us to characterize the *distributions* in the general population based on the statistical sample

We use this during model construction

Model assessment in statistics

Our main approach

Descriptive and inferential statistics

How well the model captures the statistical *distributions* in the data set; how well it allows us to characterize the *distributions* in the general population based on the statistical sample

We use this during model construction

Statistical learning and machine learning (AI, etc.)

Predictive performance !!!

Standard measures: ***accuracy*** (fraction of correctly predicted benign and pathogenic variants), ***sensitivity*** (fraction of correctly predicted pathogenic variants), ***specificity*** (fraction of correctly predicted benign variants)

Standard approach: cross-validation (train the model on a subset of the data, test on the other subset; repeat for different data partitions)

Initial analysis and approaches

Technical decisions made (data preprocessing)

- Should we filter out outliers?
- What do we do with the two “data pools” (biological replicates)?
- Should we log-transform the data?
- …(many more)

Initial analysis and approaches

Technical decisions made (data preprocessing)

- Should we filter out outliers?
- What do we do with the two “data pools” (biological replicates)?
- Should we log-transform the data?
- …(many more)

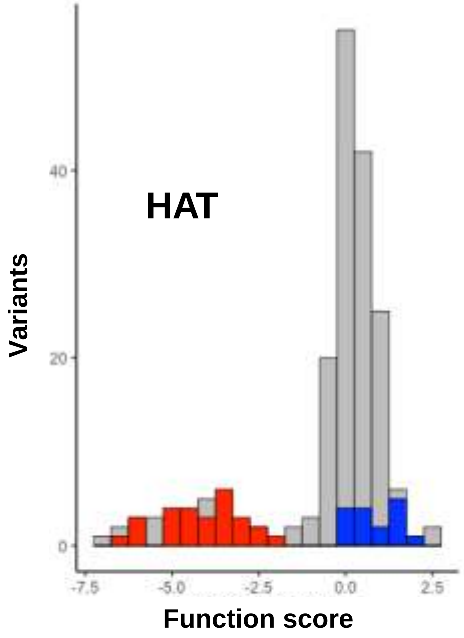
Statistical approaches considered

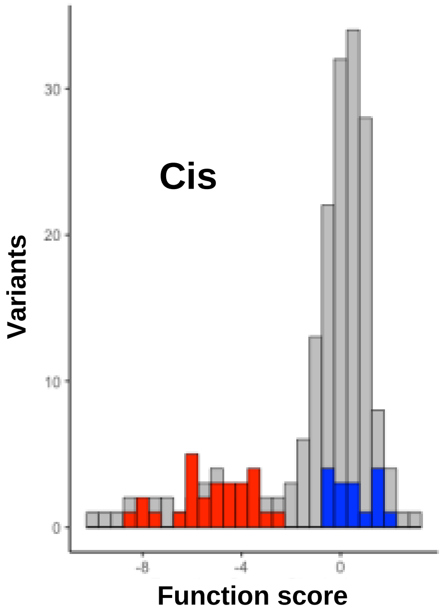
- Logistic regression
- Linear and quadratic discriminant analysis
- Mixture modeling with non-normal components
- Supervised-learning approach to mixture modeling

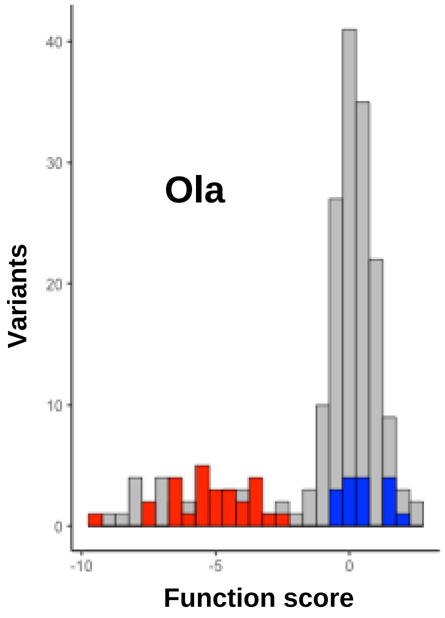
The approach that worked

N = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic
(the rest, N_u, are VUS = variants of uncertain significance)

Data distributions





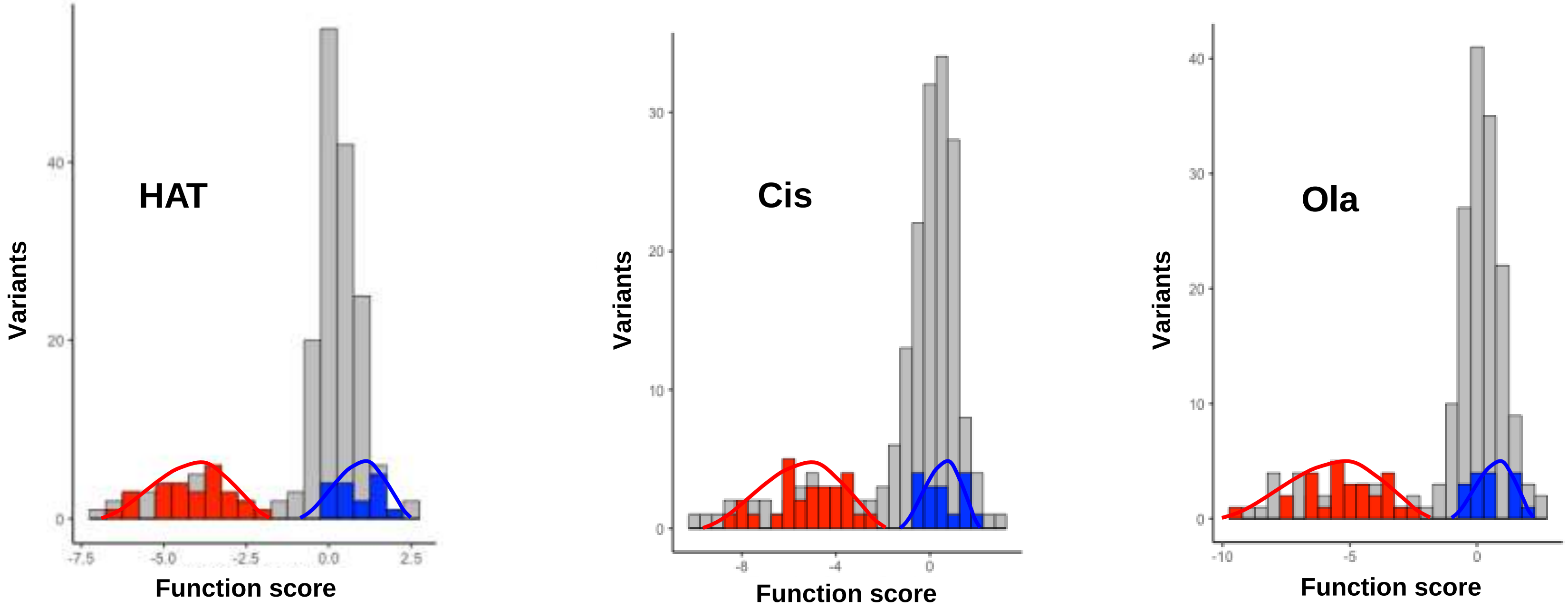


Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

The approach that worked

N = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic
(the rest, N_u, are VUS = variants of uncertain significance)

Data distributions



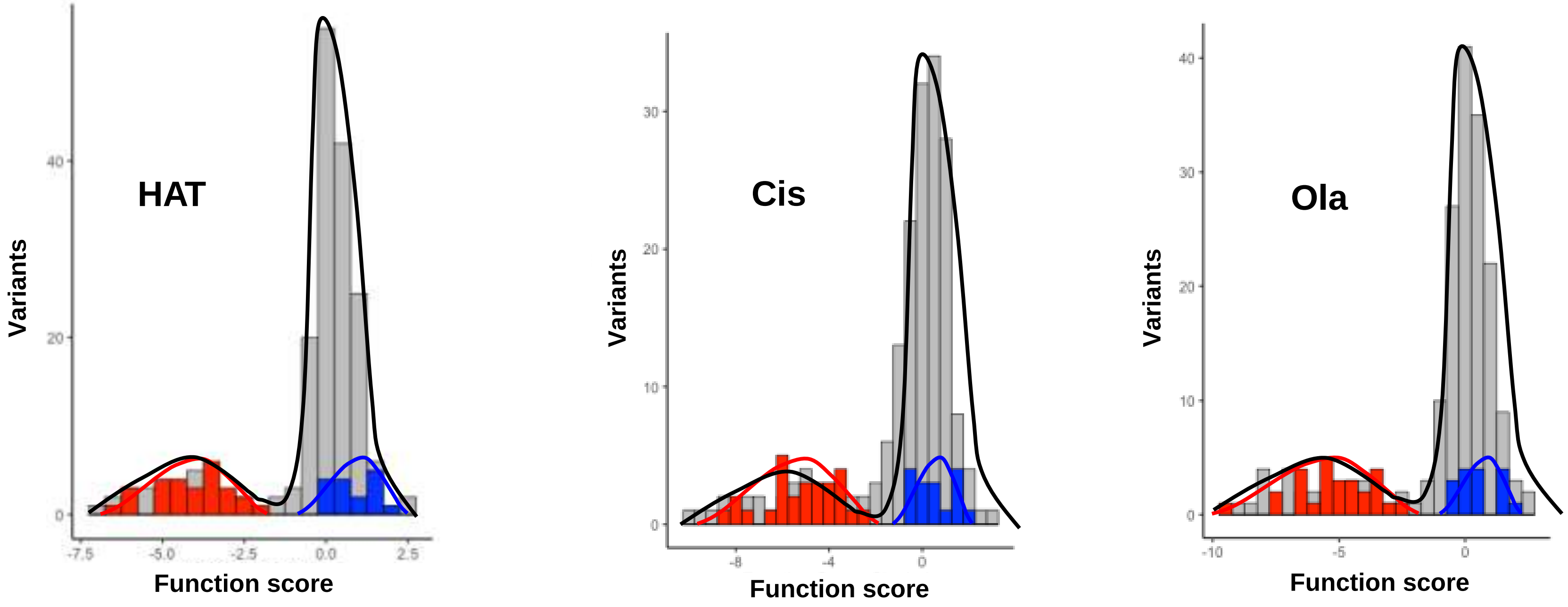
- Check the benign and pathogenic distributions for normality (fit with a bell-shaped curve)

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

The approach that worked

N = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic
(the rest, N_u, are VUS = variants of uncertain significance)

Data distributions



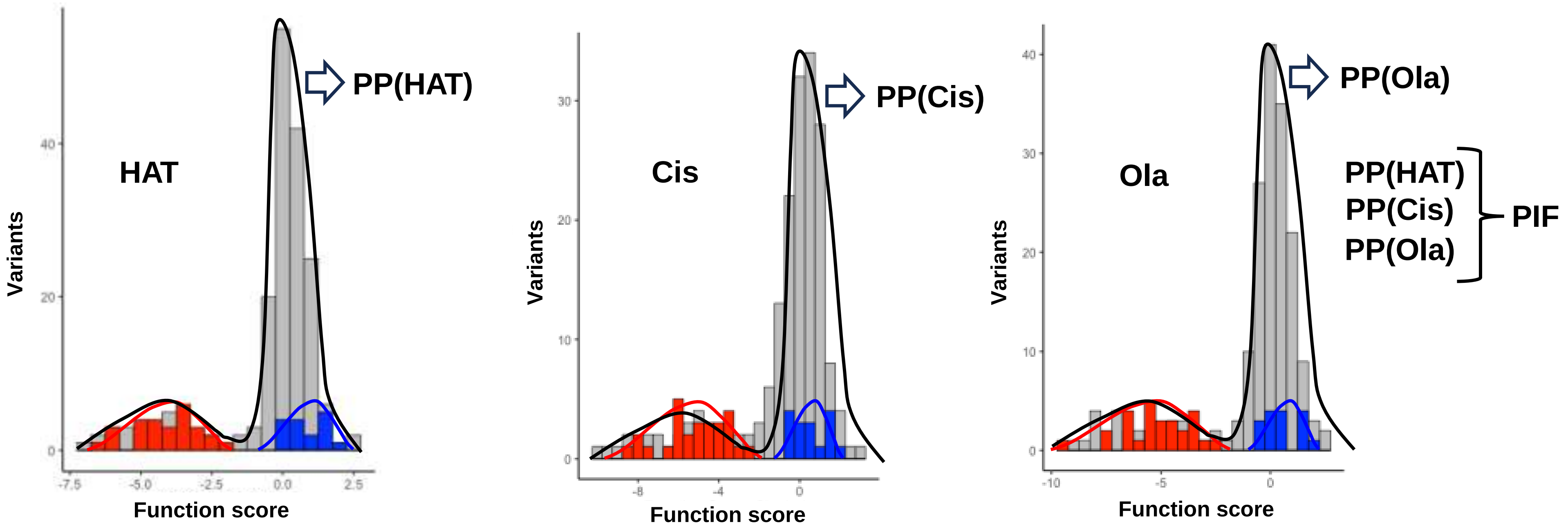
- Check the benign and pathogenic distributions for normality (fit with a bell-shaped curve)
- Fit the overall, two-peaked distribution using a combination (mixture) of benign and pathogenic distributions

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

The approach that worked

N = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic
(the rest, N_u, are VUS = variants of uncertain significance)

Data distributions



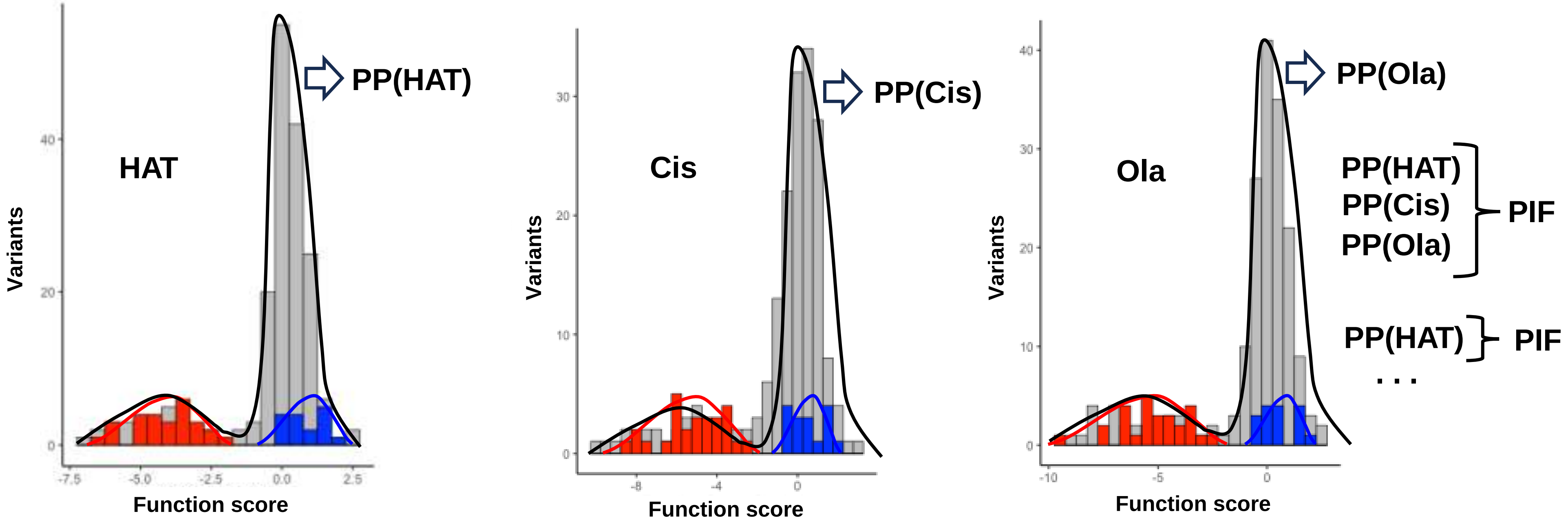
- Check the benign and pathogenic distributions for normality (fit with a bell-shaped curve)
- Fit the overall, two-peaked distribution using a combination (mixture) of benign and pathogenic distributions
- Use the parameters from the fits with some math to calculate PIFs (probabilities of impact on function)

Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

The approach that worked

N = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic
(the rest, N_u, are VUS = variants of uncertain significance)

Data distributions



- Check the benign and pathogenic distributions for normality (fit with a bell-shaped curve)
- Fit the overall, two-peaked distribution using a combination (mixture) of benign and pathogenic distributions
- Use the parameters from the fits with some math to calculate PIFs (probabilities of impact on function)
- Consider alternative models (supervised-learning and/or one input variable only, **HAT** or **Cis** or **Ola**)

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Our main methodology: mixture modeling + semi-supervised learning + empirical Bayes + heuristics

- *N* = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic (the rest, N_u, are VUS = variants of uncertain significance)

- 3 numerical variables: ***HAT***, ***Cis***, and ***Ola*** (function scores), with values for every *BRCA2* variant

- For each variable, distribution density is modeled independently as a normal mixture:

$$g(x, p, m_p, v_p, m_b, v_b) = pf(x, m_p, v_p) + (1 - p)f(x, m_b, v_b)$$

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Our main methodology: mixture modeling + semi-supervised learning + empirical Bayes + heuristics

- *N* = 223 *BRCA2* variants; N_b = 16 labeled benign and N_p = 27 labeled pathogenic (the rest, N_u, are VUS = variants of uncertain significance)

- 3 numerical variables: ***HAT***, ***Cis***, and ***Ola*** (function scores), with values for every *BRCA2* variant

- For each variable, distribution density is modeled independently as a normal mixture:

$$g(x, p, m_p, v_p, m_b, v_b) = pf(x, m_p, v_p) + (1-p)f(x, m_b, v_b)$$

- Parameters are estimated via maximum-likelihood fits (semi-supervised learning):

$$l(p, m_p, v_p, m_b, v_b \mid \boldsymbol{x}) = \prod_{i=1}^{N_p} pf\left(x_i^{(p)}, m_p, v_p\right) \times \prod_{i=1}^{N_b} (1-p) f\left(x_i^{(b)}, m_b, v_b\right) \times \prod_{i=1}^{N_u} g\left(x_i^{(u)}, p, m_p, v_p, m_b, v_b\right)$$

- Bayes formula (empirical Bayes) for probabilities of pathogenicity:

$$PP_i = \frac{pf(x_i, m_p, v_p)}{g(x_i, p, m_p, v_p, m_b, v_b)}$$

- Heuristic PIF formulas for each *BRCA2* variant:

Full: $PIF_i = PP_i(HAT) + \left(1 - PP_i(HAT)\right)PP_i(Cis)PP_i(Ola)$

Alt.: $PIF_i = PP_i(HAT)$ $\;OR\;$ $PIF_i = PP_i(Cis)$ $\;OR\;$ $PIF_i = PP_i(Ola)$

Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Validation of the computational predictions

Internal validation: K-fold cross-validation (CV) with *K* = 3, 6, 9, 43 (*K* = 43 is leave-one-out CV)

Basis: fixed benign and pathogenic PIF thresholds (**≤0.05** and **>0.99**, respectively)

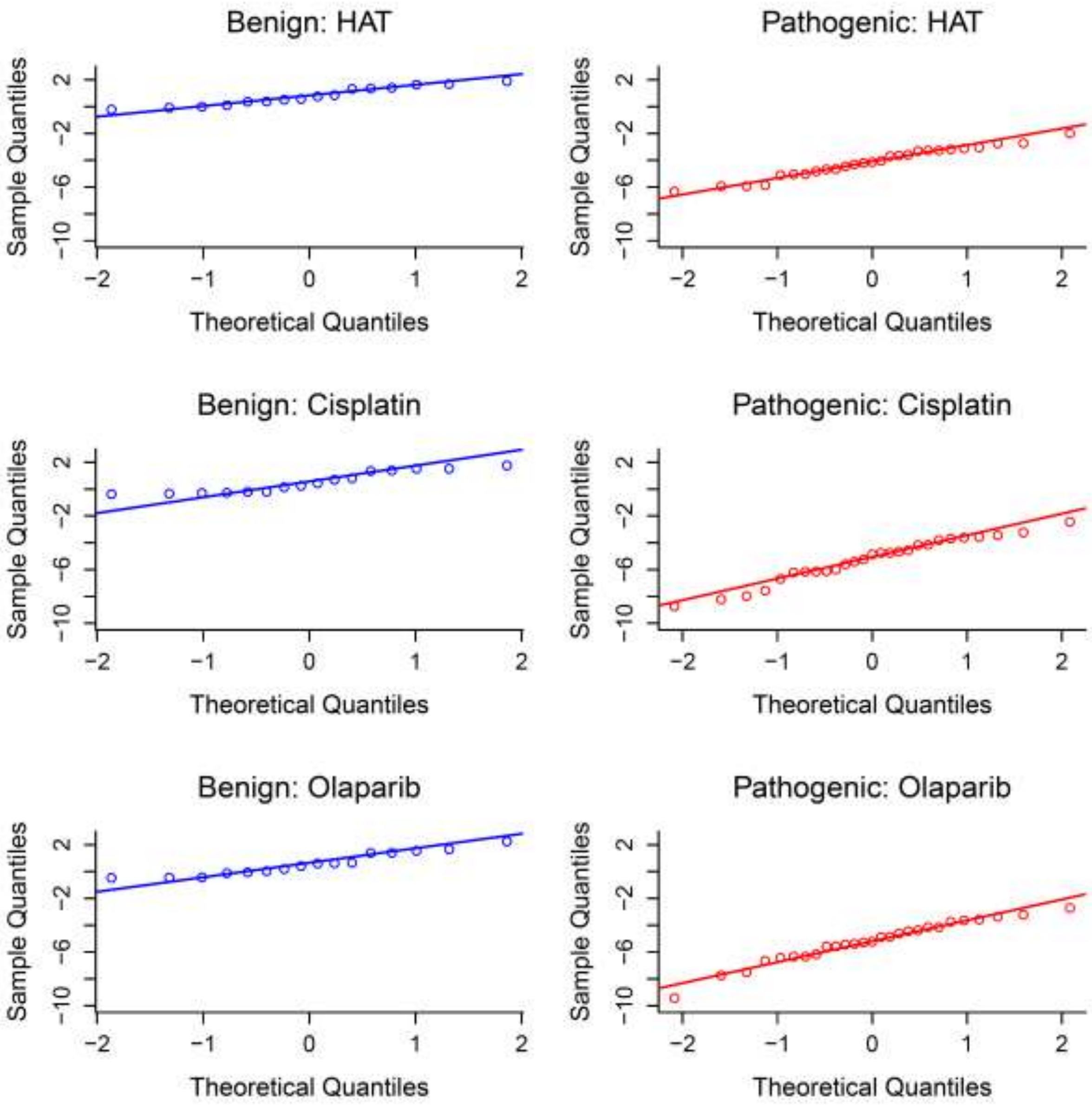
Main performance metric: ***accuracy*** (% correctly classified *BRCA2* variants, averaged across folds)

Full algorithm version (semi-supervised, HAT + Cis + Ola):
CV accuracy = 100% for all *K* values

External validation: information from diverse sources

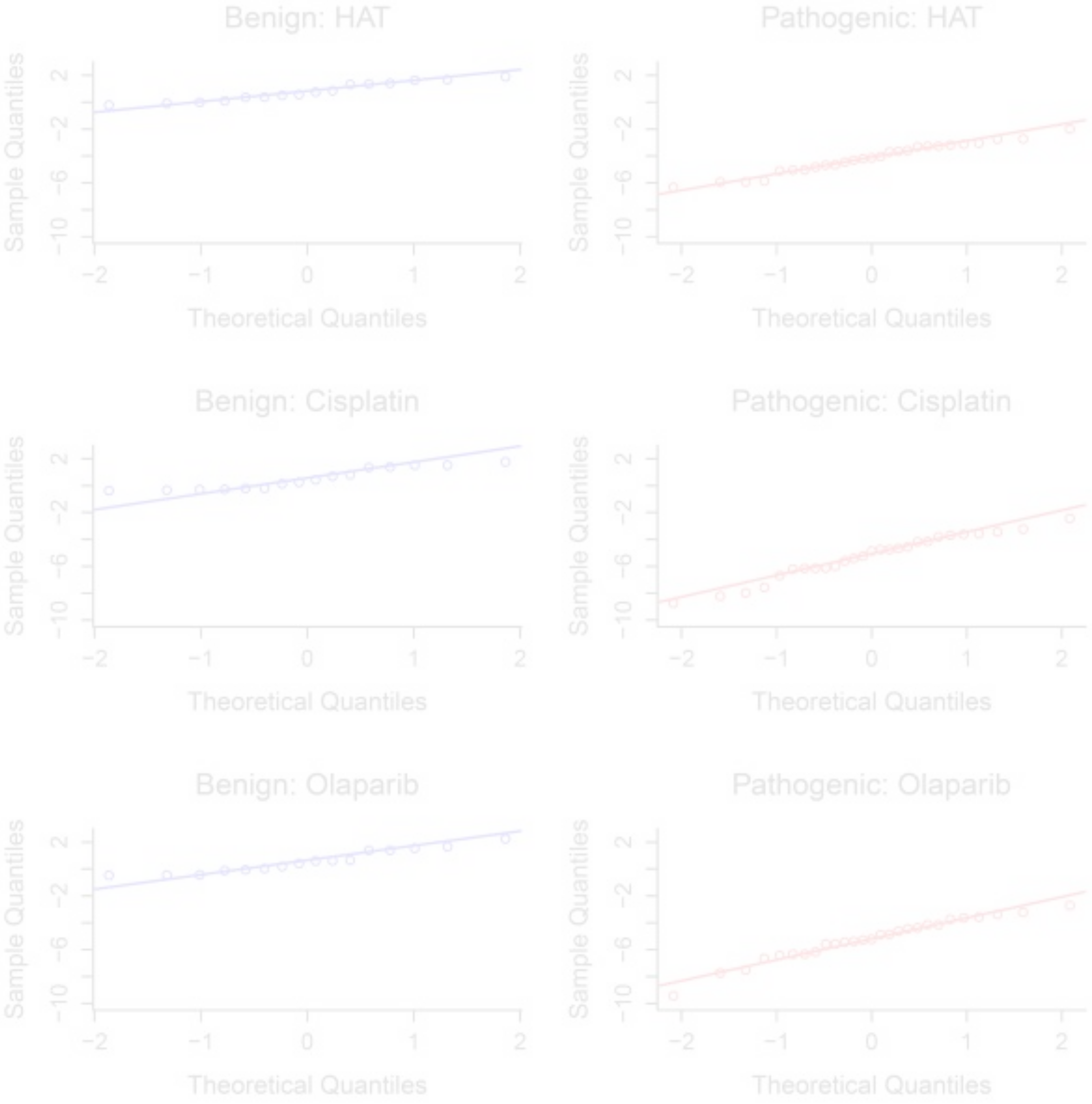
Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

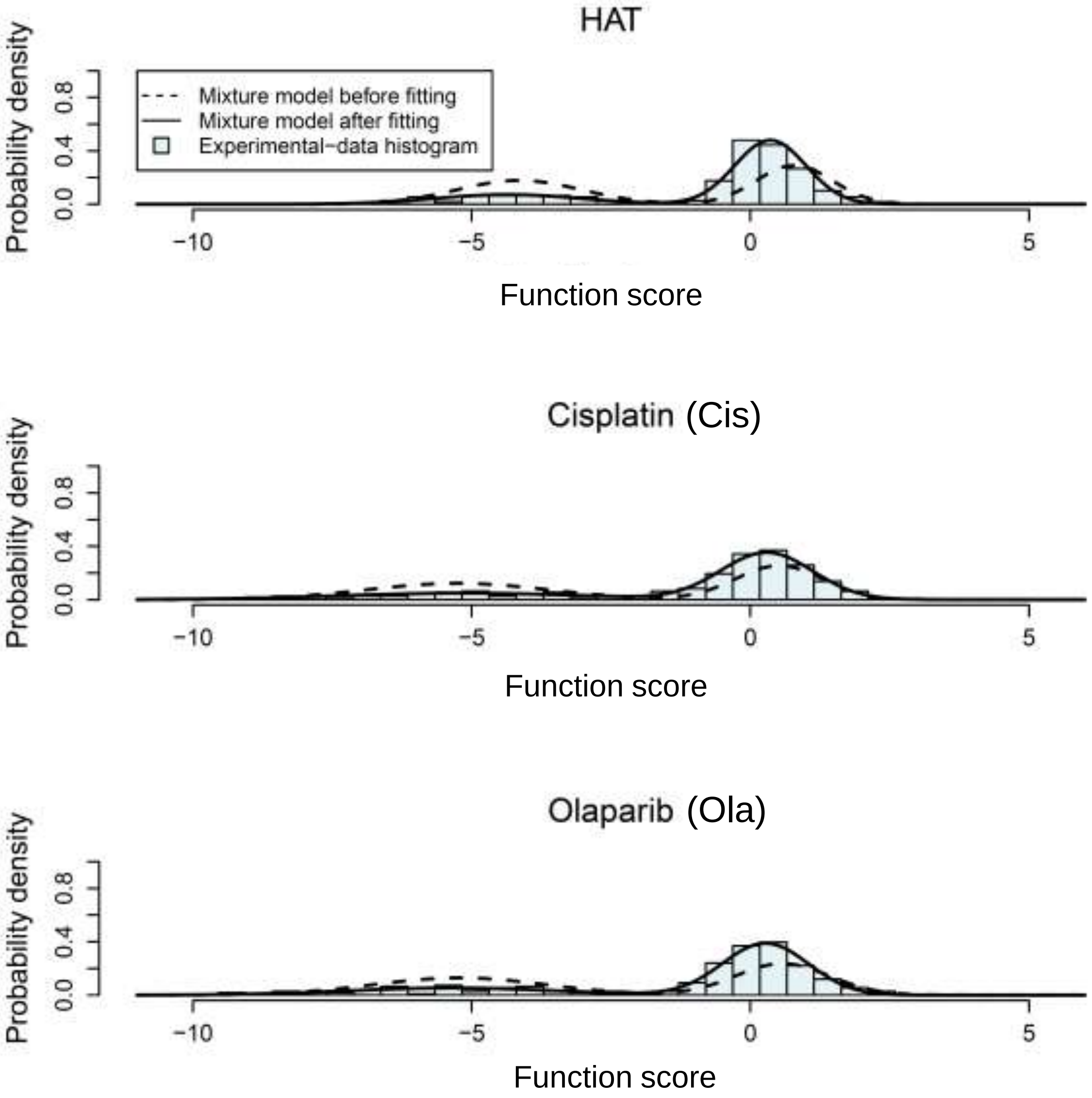
Results for data distributions: normality and fitting (mixture model)



Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

Results for data distributions: normality and fitting (mixture model)

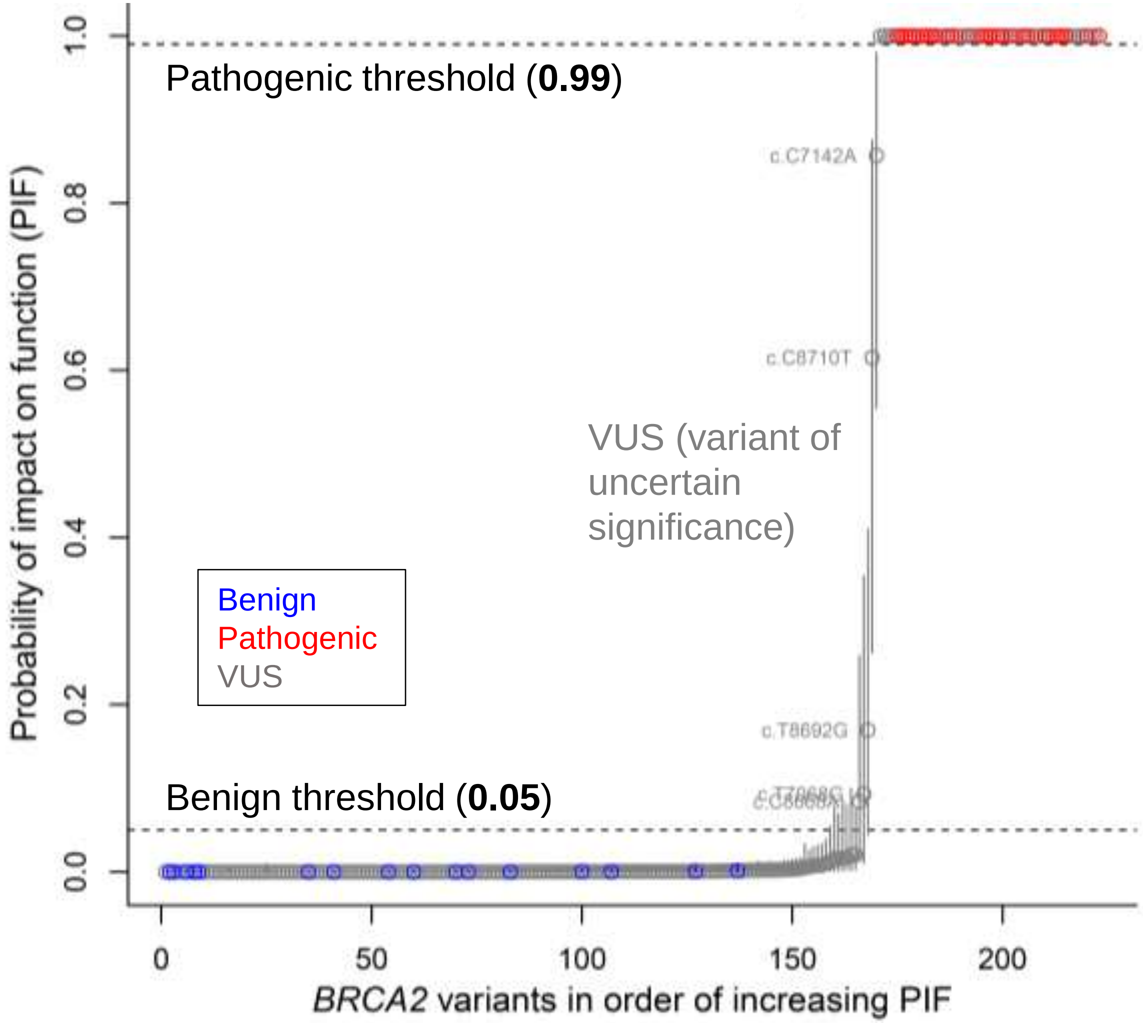




Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs)

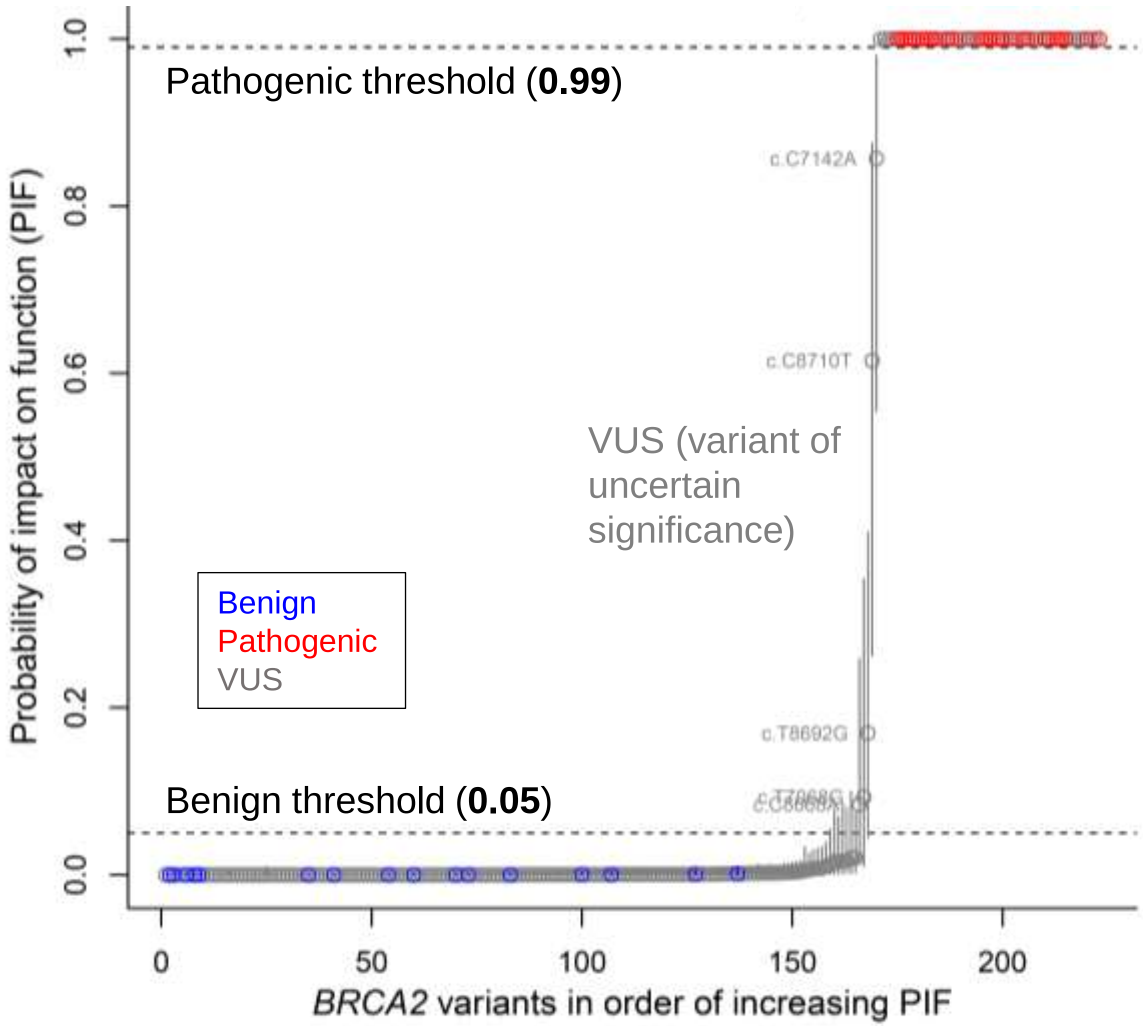
Full model: HAT + Cisplatin + Olaparib



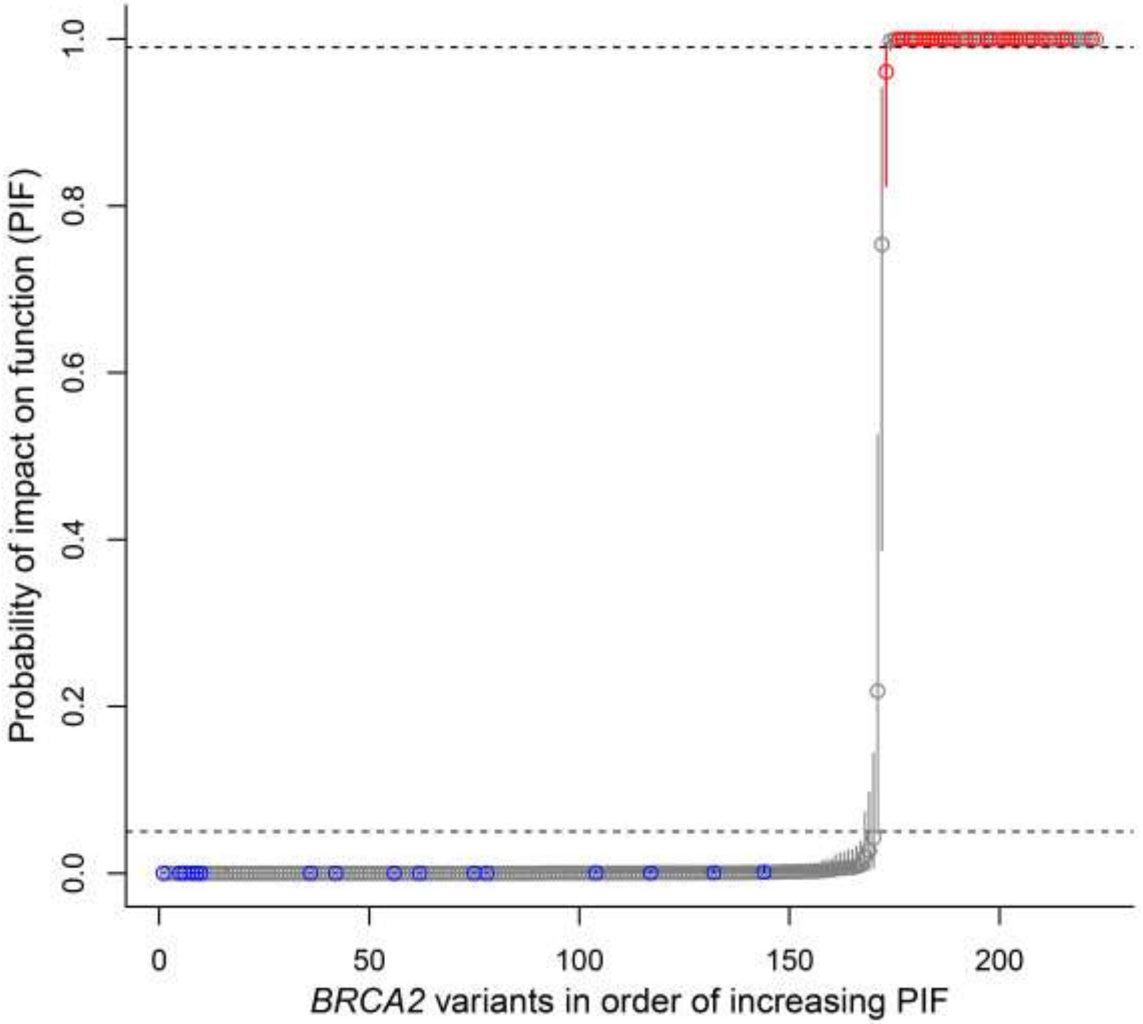
Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs)

Full model: HAT + Cisplatin + Olaparib



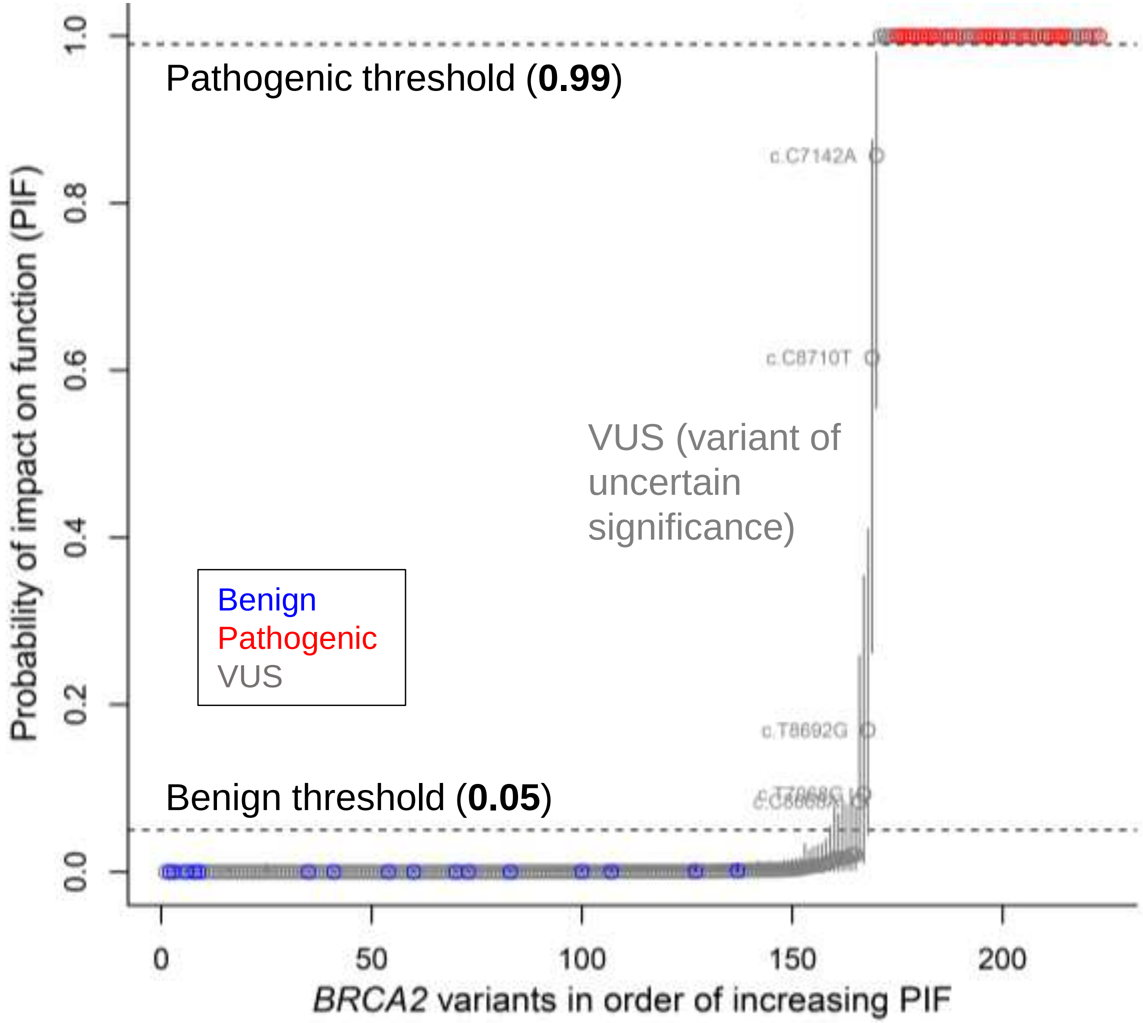
HAT



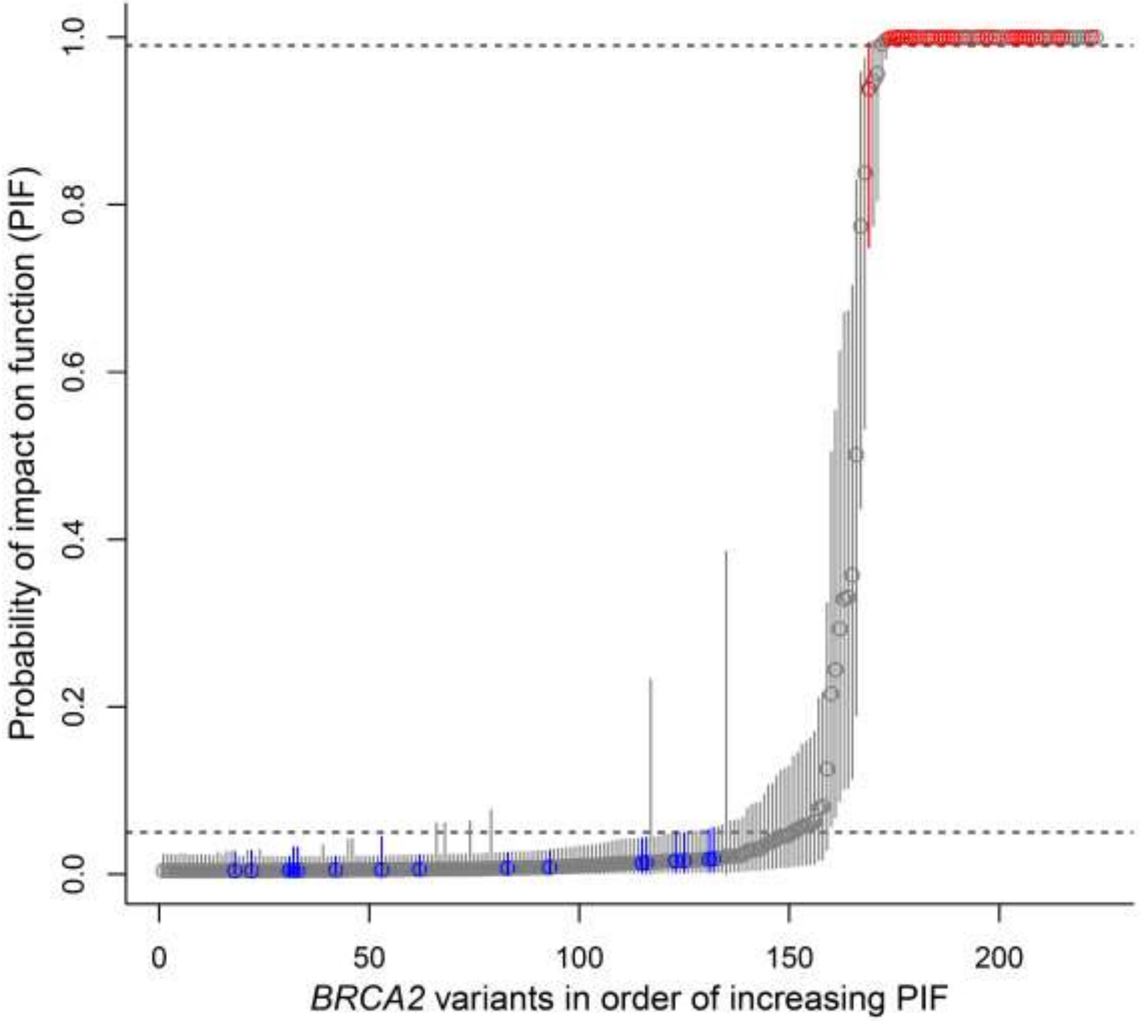
Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs)

Full model: HAT + Cisplatin + Olaparib



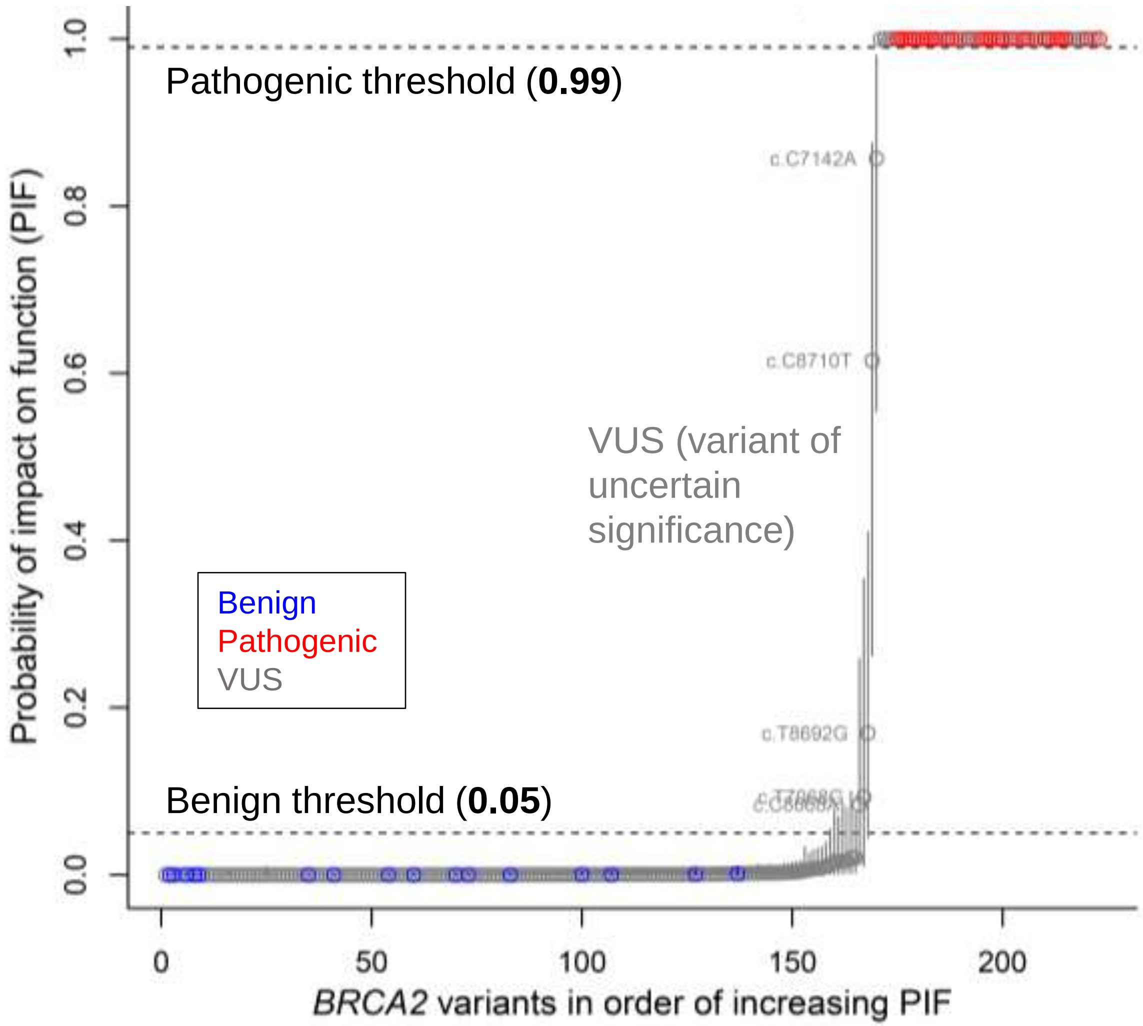
Cisplatin



Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs)

Full model: HAT + Cisplatin + Olaparib

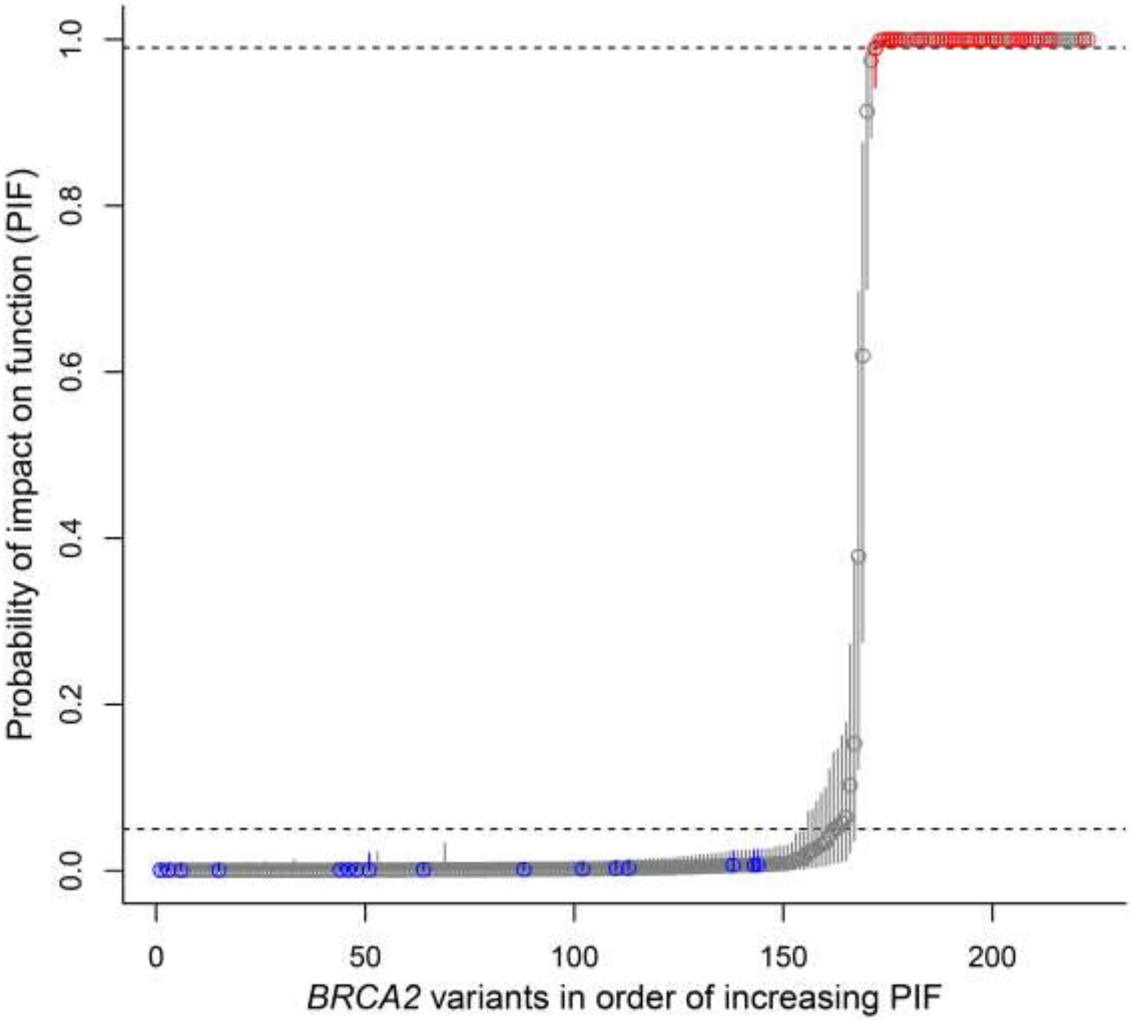


Pathogenic threshold (**0.99**)

VUS (variant of uncertain significance)

Benign threshold (**0.05**)

Olaparib

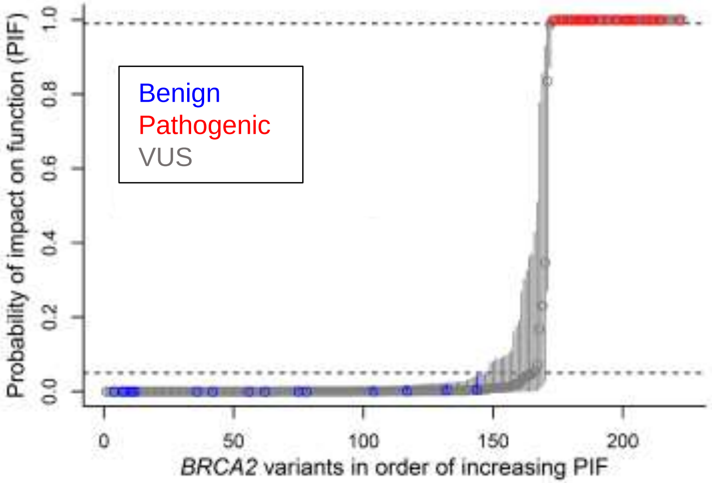


Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs): supervised learning

Computed probabilities of impact on function (PIFs): supervised learning

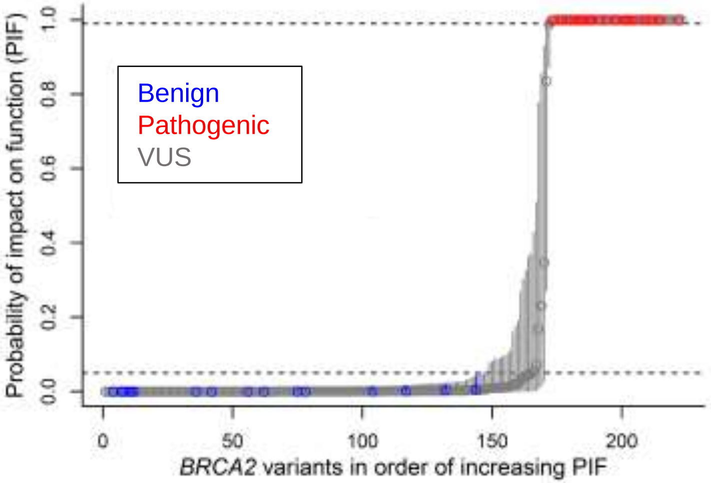
HAT

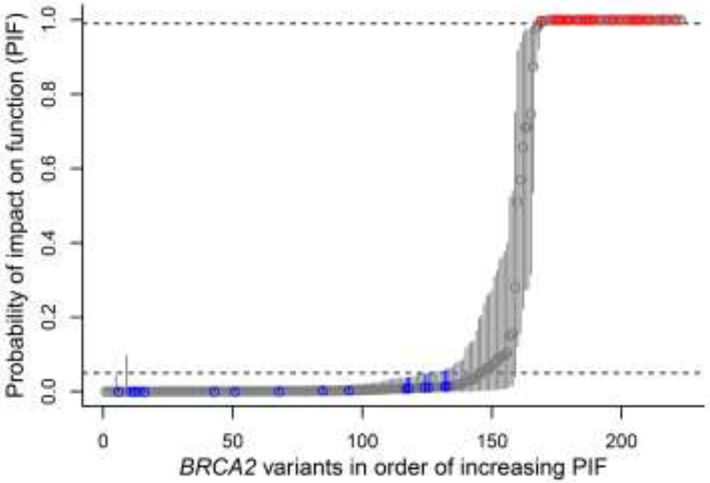


Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs): supervised learning

HAT



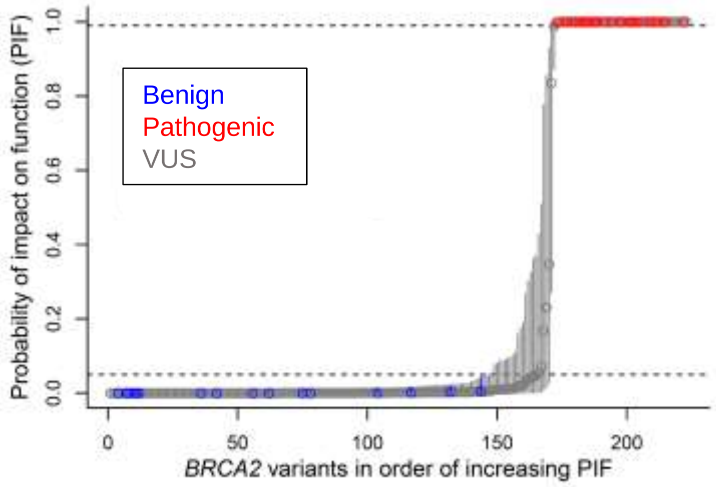


Cisplatin

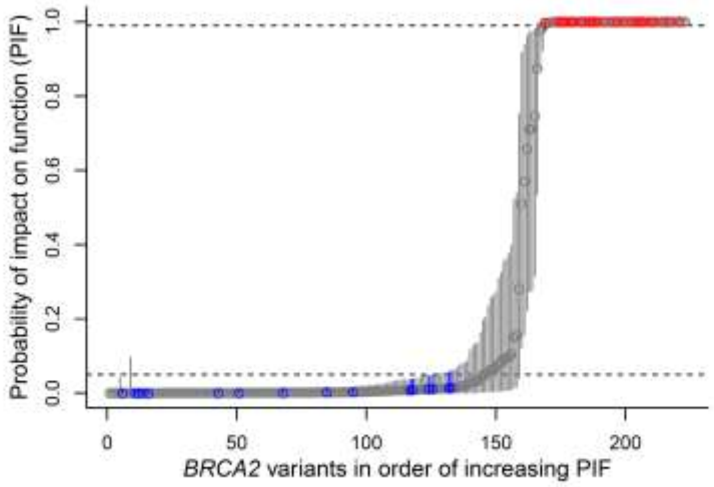
Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs): supervised learning

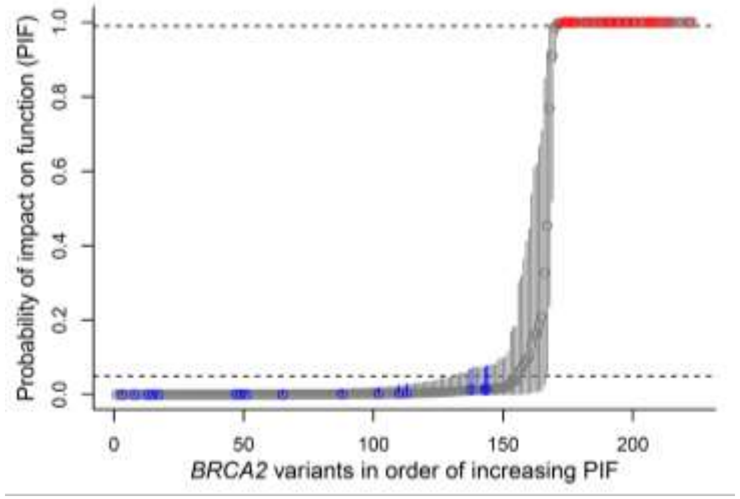
HAT



Cisplatin

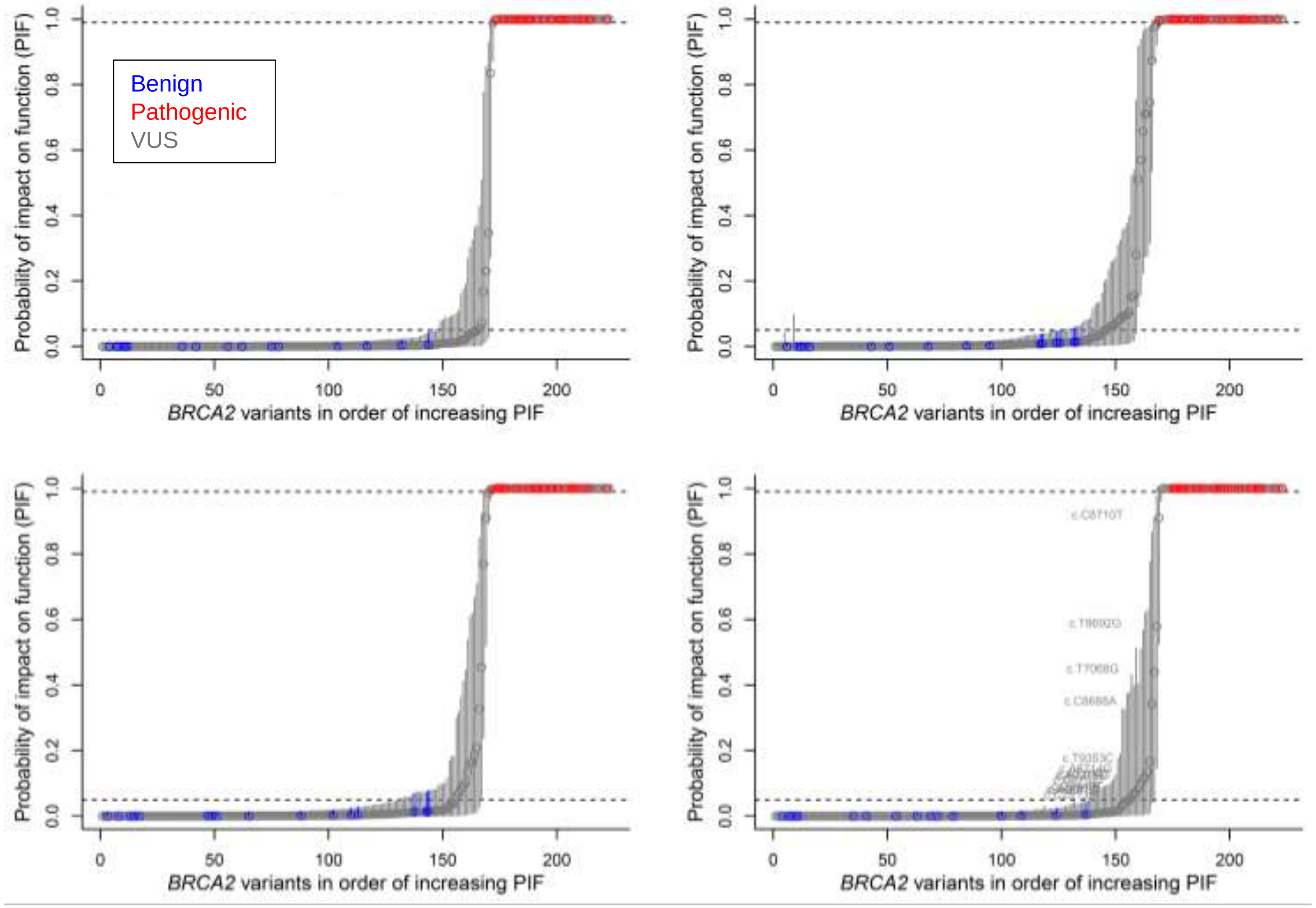


Olaparib



Biswas, Mitrophanov *et al.*, *Cell Rep Methods* 2023

Computed probabilities of impact on function (PIFs): supervised learning



HAT

Cisplatin

Olaparib

HAT + Cisplatin + Olaparib

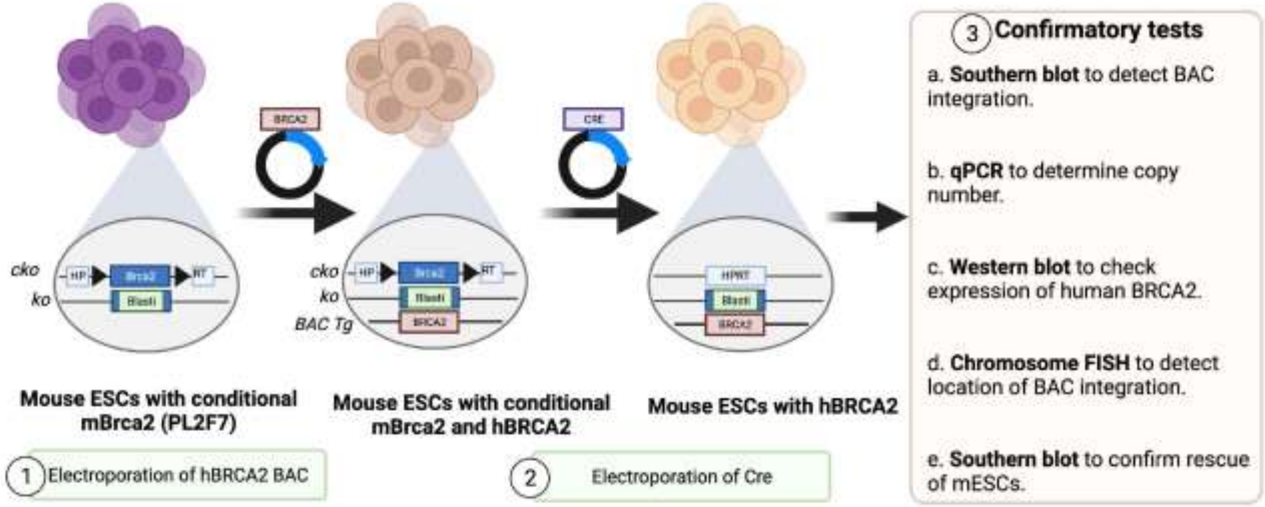
Biswas, Mitrophanov *et al*., *Cell Rep Methods* 2023

So, the full model (semi-supervised learning, all 3 variables) worked well…

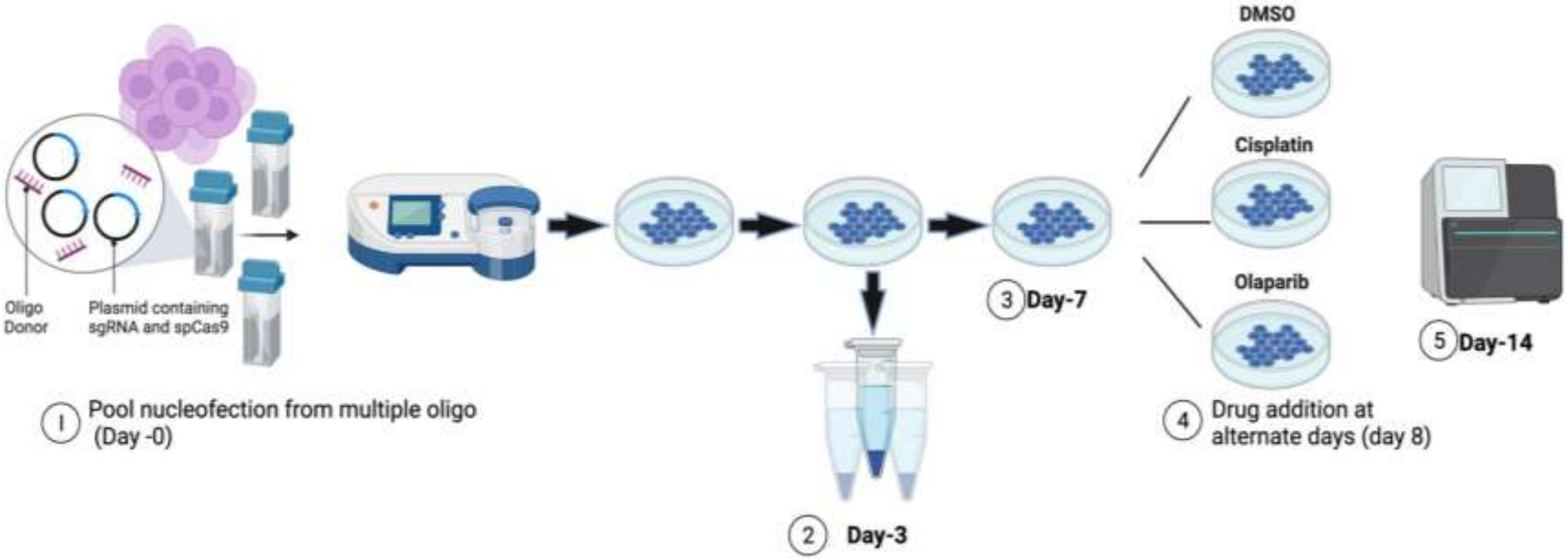
So, the full model (semi-supervised learning, all 3 variables) worked well…

But a new experimental technology required a different statistical approach

The new data (N = 599): an advanced methodology

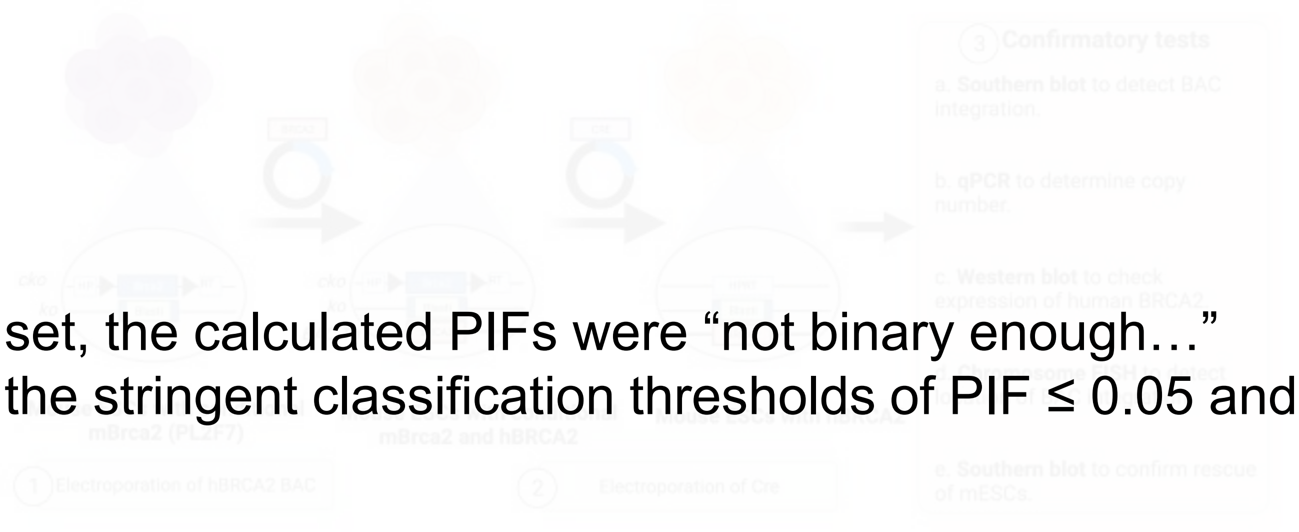

BRCA2
CRE
cko
ko
HP
Brca2
RT
Blasti
BAC Tg
BRCA2
HPRT
Mouse ESCs with conditional mBrca2 (PL2F7)
Mouse ESCs with conditional mBrca2 and hBRCA2
Mouse ESCs with hBRCA2
1 Electroporation of hBRCA2 BAC
2 Electroporation of Cre
3 Confirmatory tests
a. Southern blot to detect BAC integration.
b. qPCR to determine copy number.
c. Western blot to check expression of human BRCA2.
d. Chromosome FISH to detect location of BAC integration.
e. Southern blot to confirm rescue of mESCs.

Essential aspect: CRISPR technology!!


Oligo Donor
Plasmid containing sgRNA and spCas9
1 Pool nucleofection from multiple oligo (Day -0)
2 Day-3
3 Day-7
DMSO
Cisplatin
Olaparib
4 Drug addition at alternate days (day 8)
5 Day-14

Sahu, Sullivan, Mitrophanov *et al.*, *PLOS Genet* 2023

The new data (N = 599): an advanced methodology



mBrca2 (PL2F7)
mBrca2 and hBRCA2
Mouse ESCs with hBRCA2
1 Electroporation of hBRCA2 BAC
2 Electroporation of Cre
3 Confirmatory tests
a. Southern blot to detect BAC integration.
b. qPCR to determine copy number.
c. Western blot to check expression of human BRCA2
d. Chromosome FISH to detect
e. Southern blot to confirm rescue of mESCs.

Essential aspect: CRISPR technology!!

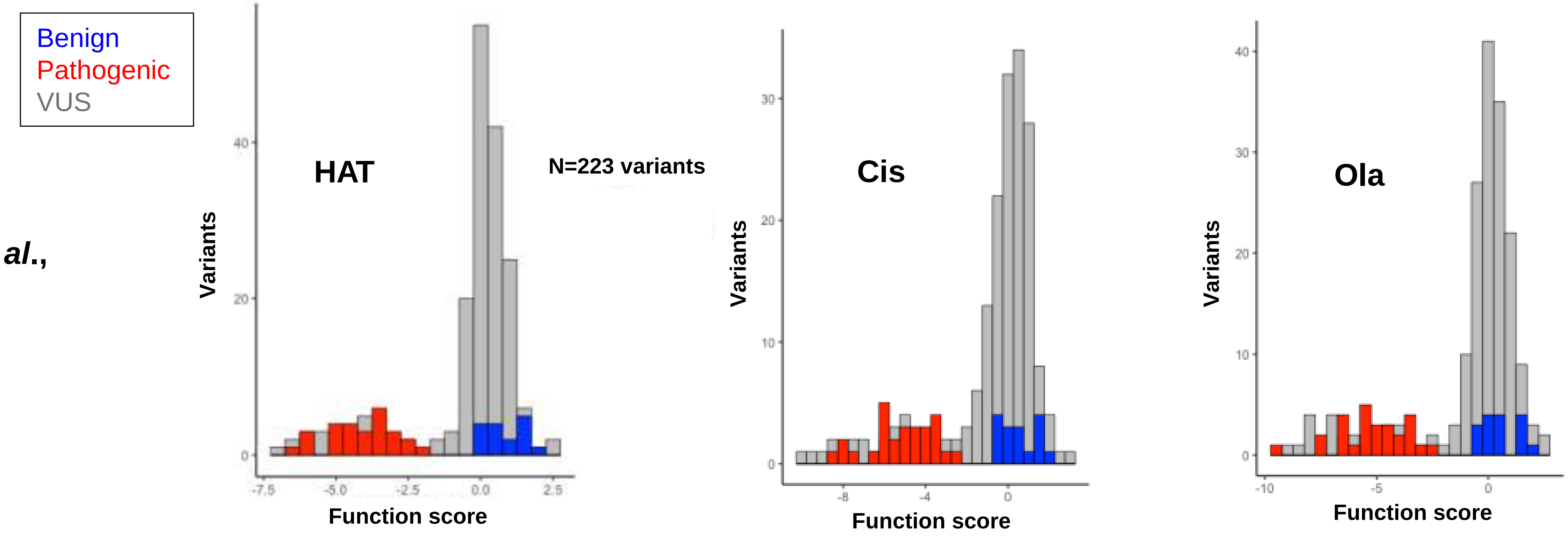
On the **new** data set, the calculated PIFs were “not binary enough…”
(i.e., not meeting the stringent classification thresholds of PIF ≤ 0.05 and PIF > 0.99)

But WHY ??

Oligo Donor
Plasmid containing sgRNA and spCas9
1 Pool nucleofection from multiple oligo (Day -0)
2 Day-3
3 Day-7
DMSO
Cisplatin
Olaparib
4 Drug addition at alternate days (day 8)
5 Day-14

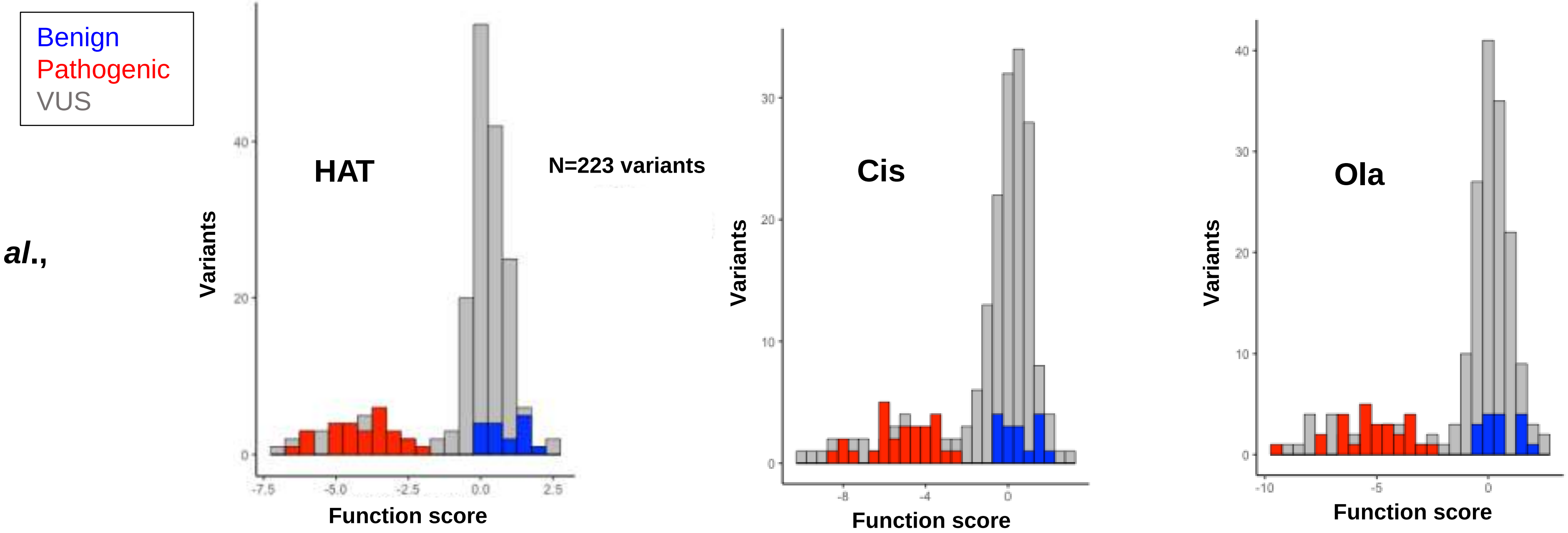
Apparent reason: insufficient distribution separation

Biswas, Mitrophanov *et al*.,
***Cell Rep Methods* 2023**

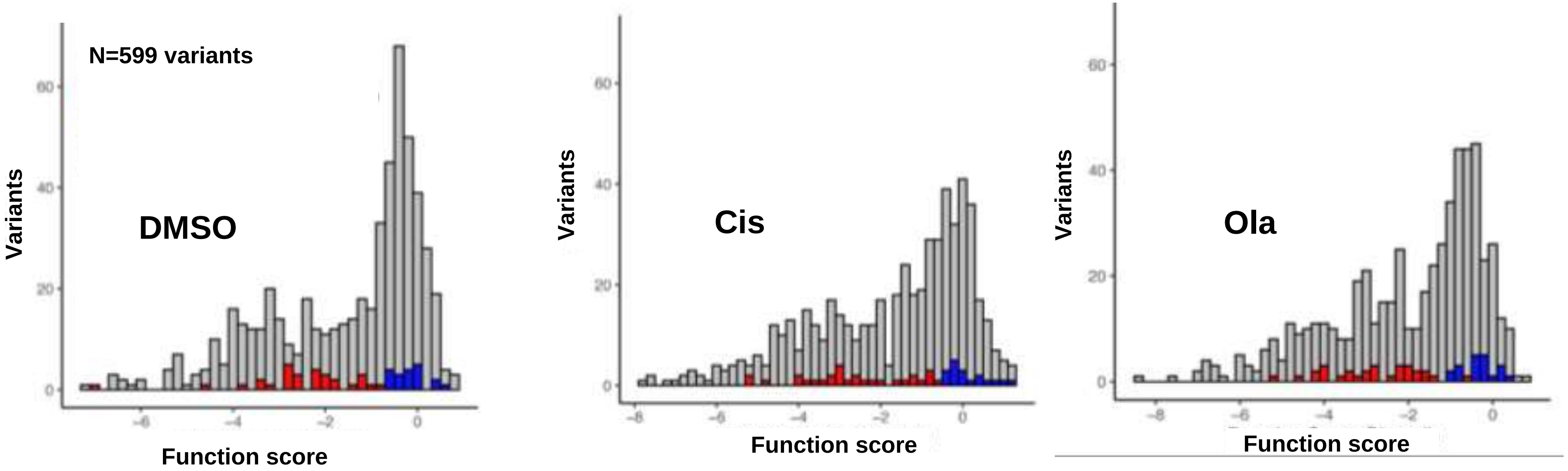


Apparent reason: insufficient distribution separation

Biswas, Mitrophanov *et al*.,
***Cell Rep Methods* 2023**



Sahu, Sullivan, Mitrophanov *et al*.,
***PLOS Genet* 2023**



The probit regression model

Just like logistic (=logit) regression, only with a different link function instead of log-odds

The standard supervised-learning approach !!

Sahu, Sullivan, Mitrophanov *et al.*, *PLOS Genet* 2023

The probit regression model

The standard supervised-learning approach !!

- N = 599 *BRCA2* variants; N_b = 21 labeled benign and N_p = 29 labeled pathogenic (the rest, N_u, are VUS = variants of uncertain significance)

- 3 numerical variables: ***DMSO***, ***Cis***, and ***Ola*** (function scores), with values for every *BRCA2* variant

- PIF formulas for each *BRCA2* variant from classic probit regression:

Full: $$\text{probit}(PIF_i) = b_0 + b_1 DMSO_i + b_2 Cis_i + b_3 Ola_i$$

Sahu, Sullivan, Mitrophanov *et al.*, *PLOS Genet* 2023

The probit regression model

The standard supervised-learning approach !!

- N = 599 *BRCA2* variants; N_b = 21 labeled benign and N_p = 29 labeled pathogenic (the rest, N_u, are VUS = variants of uncertain significance)

- 3 numerical variables: ***DMSO***, ***Cis***, and ***Ola*** (function scores), with values for every *BRCA2* variant

- PIF formulas for each *BRCA2* variant from classic probit regression:

Full: $\mathrm{probit}(PIF_i) = b_0 + b_1 DMSO_i + b_2 Cis_i + b_3 Ola_i$

Alternatives: $\mathrm{probit}(PIF_i) = b_0 + b_1 DMSO_i$

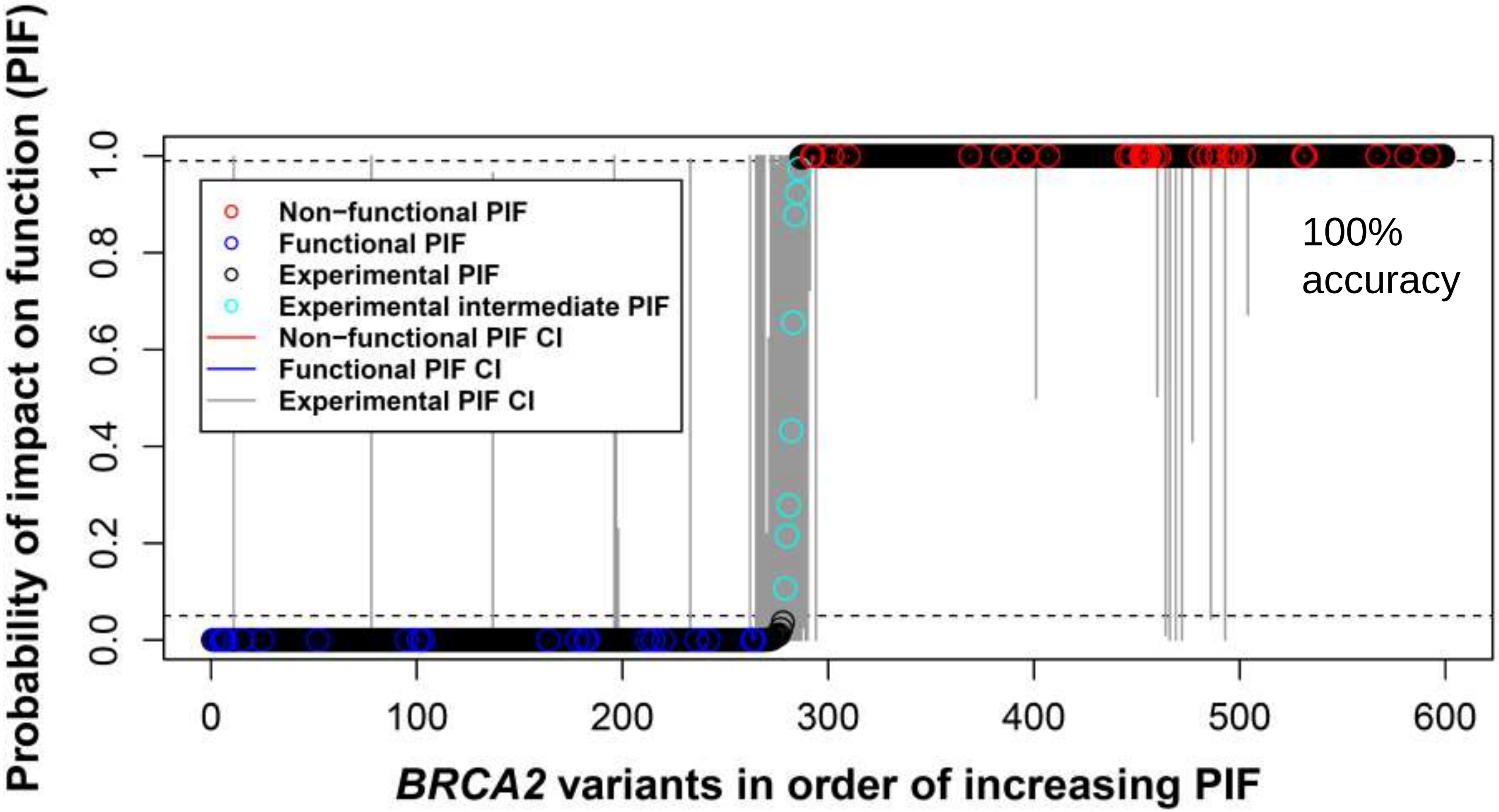
$\mathrm{probit}(PIF_i) = b_0 + b_1 Cis_i$

$\mathrm{probit}(PIF_i) = b_0 + b_1 Ola_i$

Validation approaches: cross-validation (accuracy, sensitivity, specificity), information from diverse sources

Sahu, Sullivan, Mitrophanov *et al.*, *PLOS Genet* 2023

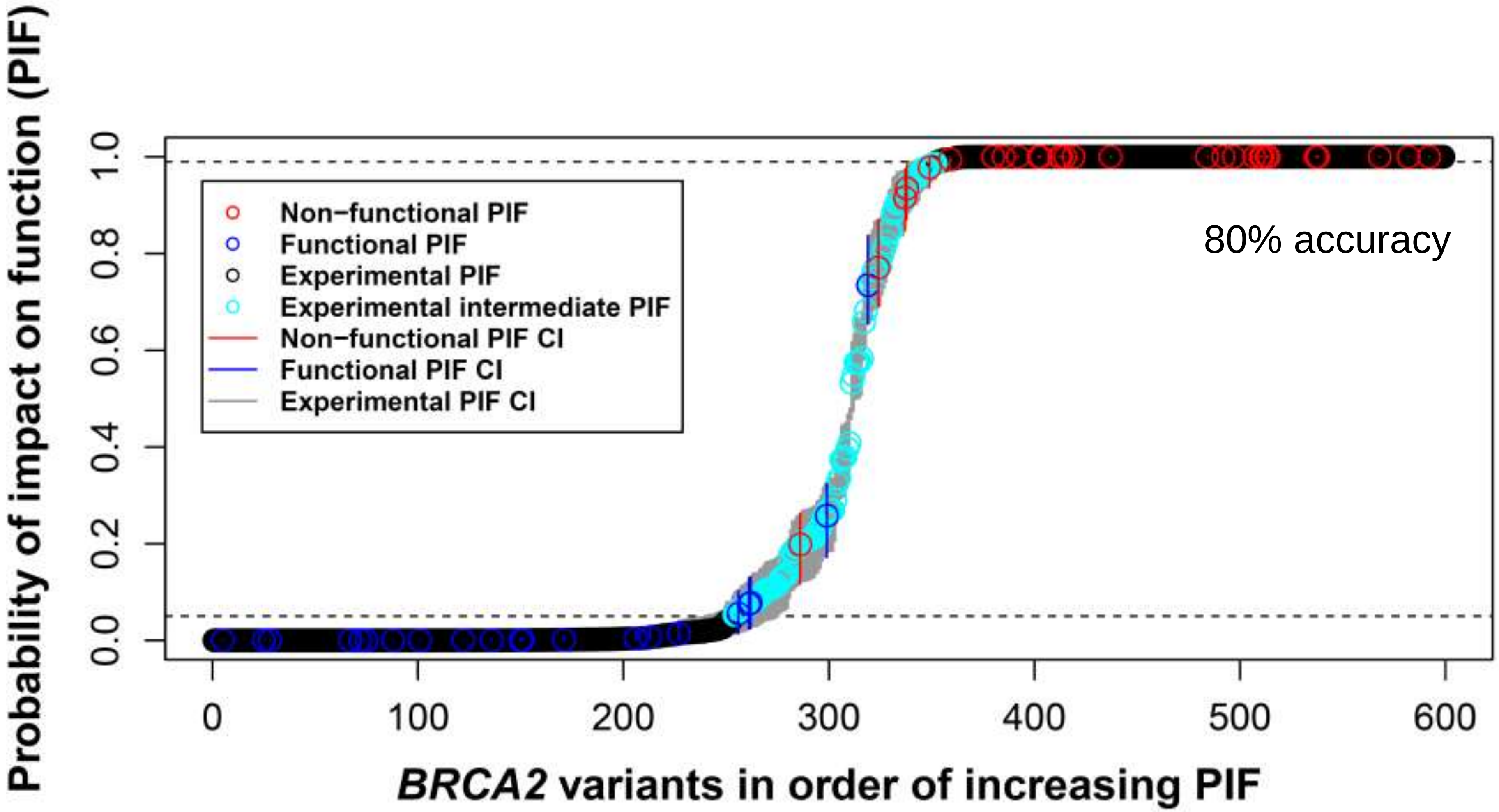
PIFs for the full (DMSO + Cisplatin + Olaparib) model



100%
accuracy

Sahu, Sullivan, Mitrophanov *et al.*, *PLOS Genet* 2023

PIFs for the DMSO model

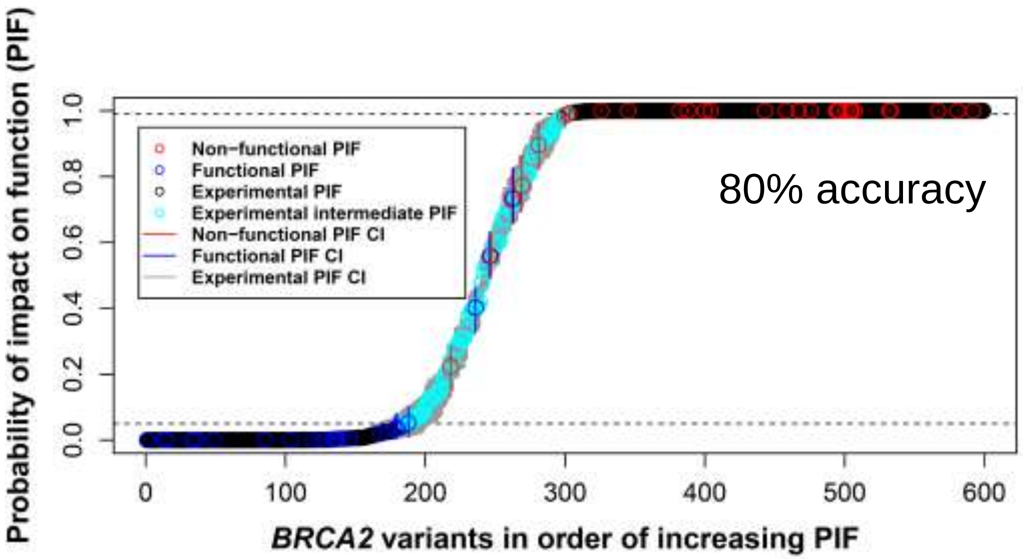


80% accuracy

Sahu, Sullivan, Mitrophanov *et al*., *PLOS Genet* 2023

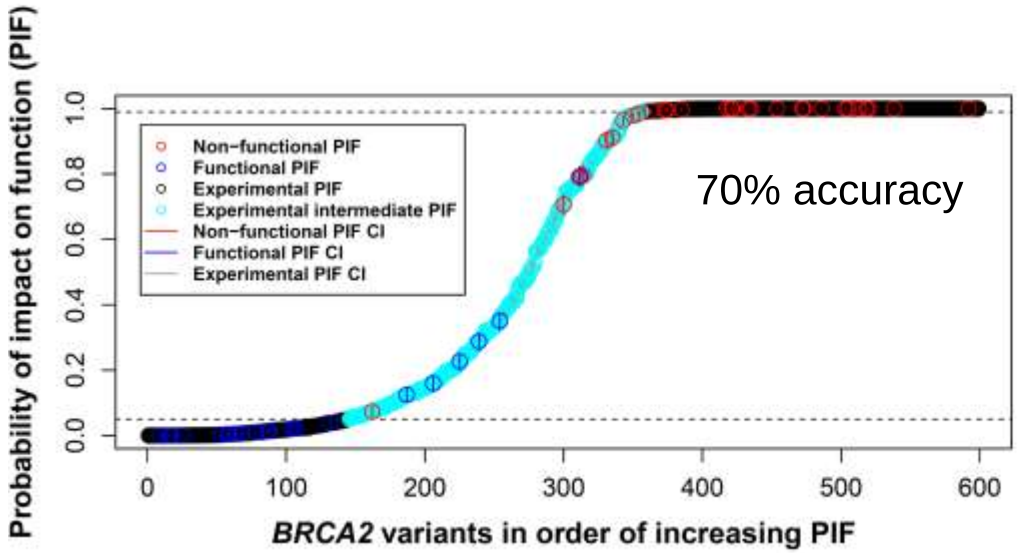
PIFs for the Cisplatin and Olaparib models

Cisplatin



80% accuracy

Olaparib



70% accuracy

Sahu, Sullivan, Mitrophanov *et al*., *PLOS Genet* 2023

Full probit-regression model: cross-validation results

Accuracy (K = 5, 10, 50): **94**%, **92**%, and **92**%

<Sensitivity, Specificity> (K = 5, 10, 50): <**92.5**%, **93.3**%>
<**91.7**%, **87.5**%>
<**93.1**%, **90.5**%>

Could we do better? Great question for future research!!

Sahu, Sullivan, Mitrophanov *et al*., *PLOS Genet* 2023

Summary

- Developed and validated new statistical approaches for computing probabilities of impact on function (PIF) for *BRCA2* variants using functional-assay data

Summary

- Developed and validated new statistical approaches for computing probabilities of impact on function (PIF) for *BRCA2* variants using functional-assay data

- Predicted the pathogenicity of hundreds of *BRCA2* variants of uncertain significance

Summary

- Developed and validated new statistical approaches for computing probabilities of impact on function (PIF) for *BRCA2* variants using functional-assay data

- Predicted the pathogenicity of hundreds of *BRCA2* variants of uncertain significance

- Performance of a particular PIF-calculation method strongly depends on the statistical distributions of the data

Summary

- Developed and validated new statistical approaches for computing probabilities of impact on function (PIF) for *BRCA2* variants using functional-assay data

- Predicted the pathogenicity of hundreds of *BRCA2* variants of uncertain significance

- Performance of a particular PIF-calculation method strongly depends on the statistical distributions of the data

- Accurate and robust out-of-distribution analysis (i.e., broad generalization capability) appears to be a challenge in PIF calculation from functional-assay data

Acknowledgements

- Tyler Malys, PhD (DMS/FNLCR)
- Duncan Donohue, PhD (ABCS/DMS/FNLCR)
- Shyam Sharan, PhD (NCI)
- Kajal Biswas, PhD (NCI)
- Sounak Sahu, PhD (NCI)

QUESTIONS?