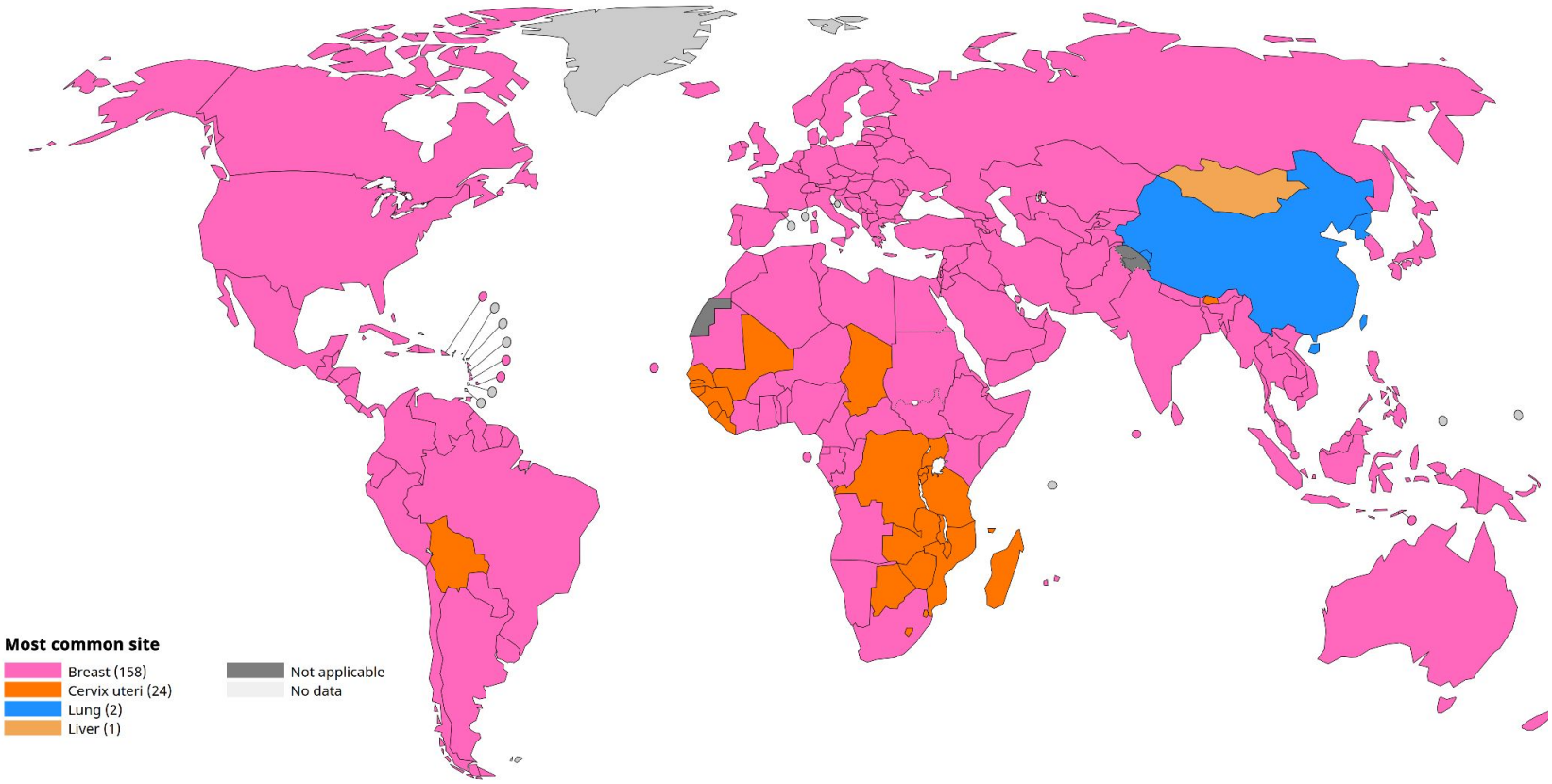
Precision Medicine for All: Using Tidymodels to Validate Breast Cancer PRS in Brazil

Flávia E. Rius, PhD

Data Scientist - Mendelics

Most common site per country, Absolute numbers, Incidence, Females, in 2022 (excl. NMSC)



Cancer TODAY | IARC
https://gco.iarc.who.int/today
Data version: Globocan 2022 (version 1.1) - 08.02.2024

International Agency for Research on Cancer
World Health Organization

Main Risk Factors

- Sex ♀

Main Risk Factors

- Sex ♀
- Age ↑

Main Risk Factors

- Sex ♀
- Age ⬆
- Lifestyle 🍷

Main Risk Factors

- Sex ♀
- Age ⬆
- Lifestyle 🍷
- Hormonal

Main Risk Factors

- Sex ♀
- Age ⬆
- Lifestyle 🍷
- Hormonal
- Family history

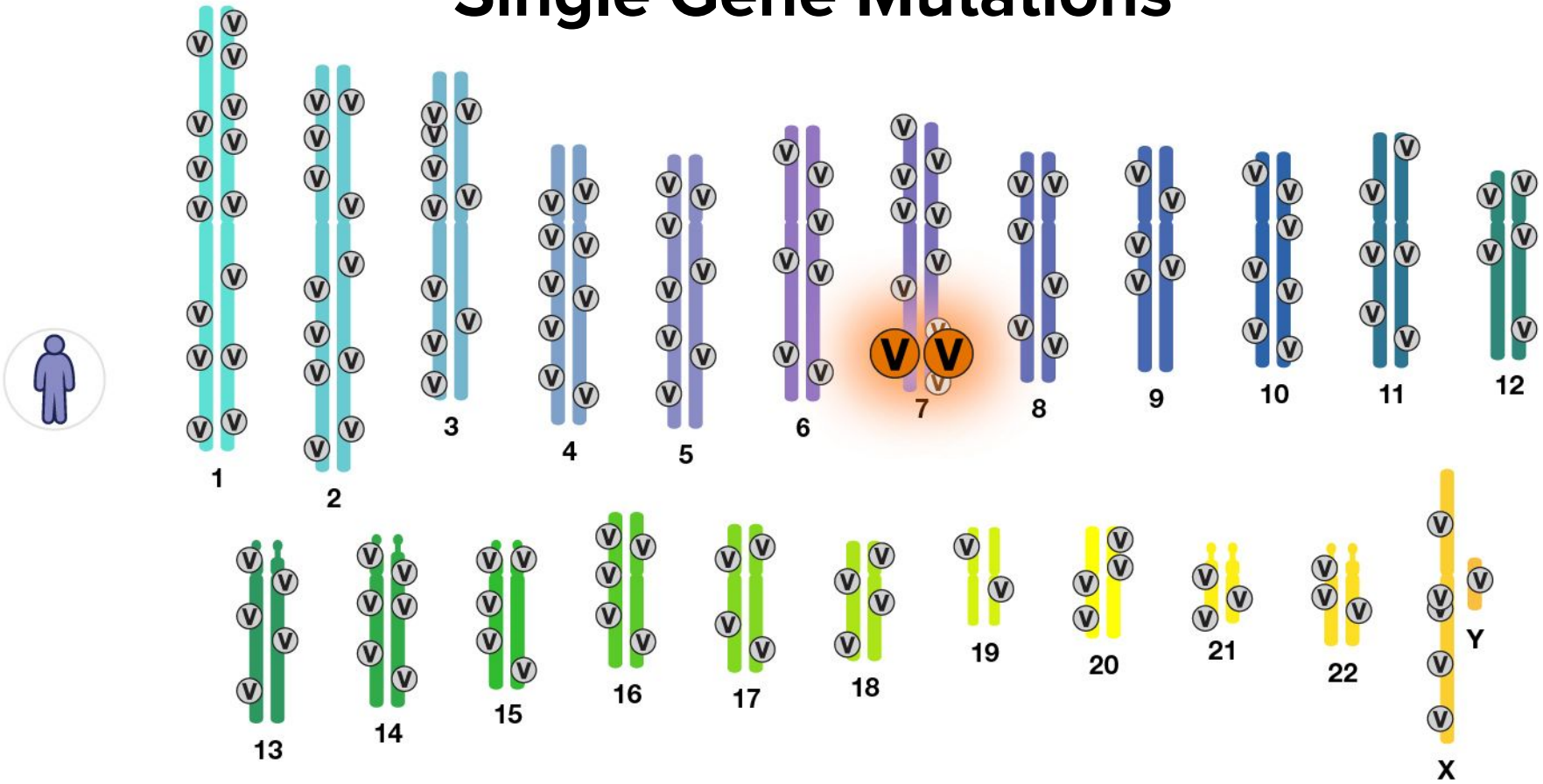
Main Risk Factors

- Sex ♀
- Age ⬆
- Lifestyle 🍷
- Hormonal
- Family history
- Genetics 🧬

Genetics of Breast Cancer

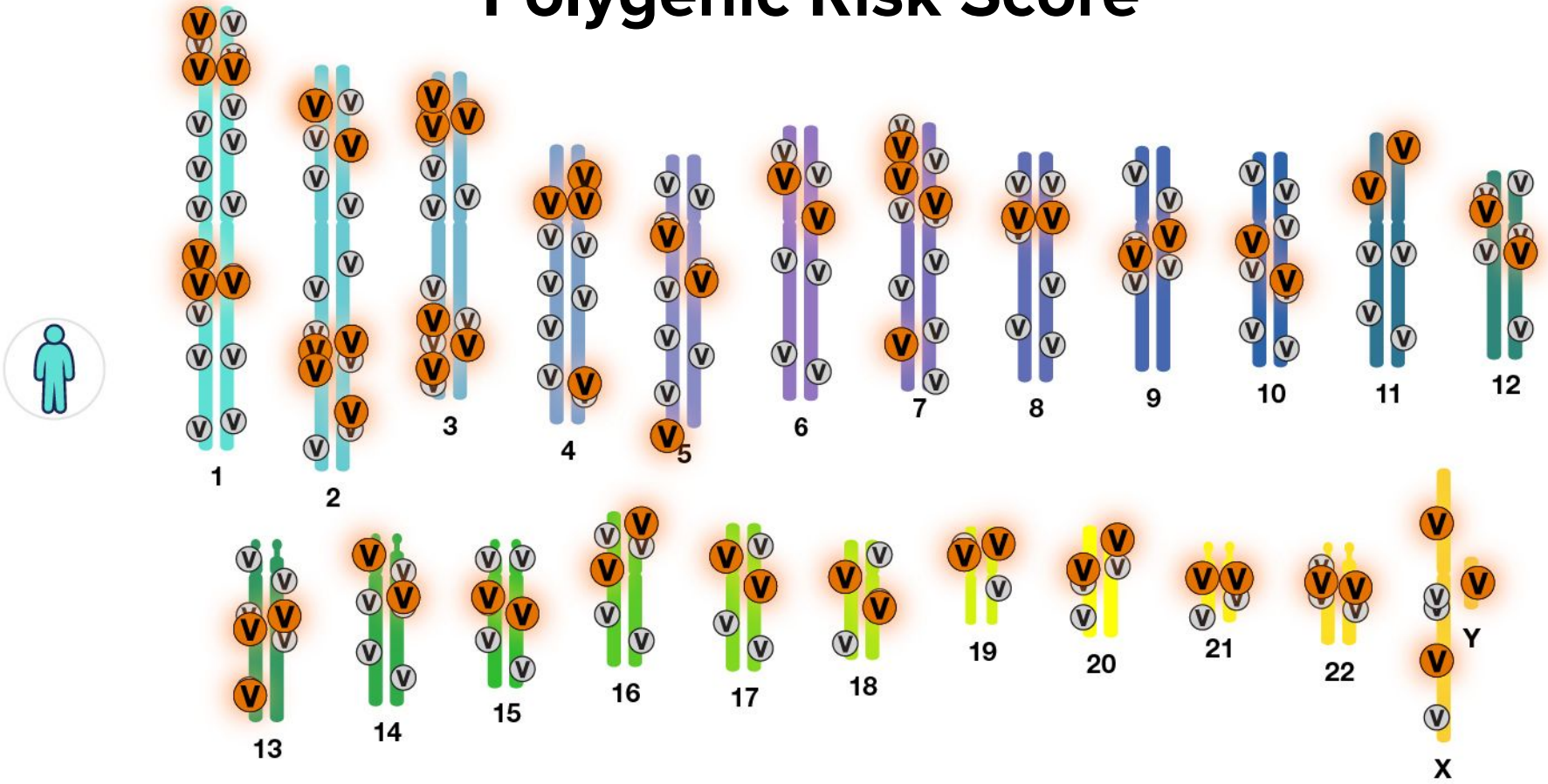
- *BRCA1* mutation

Single Gene Mutations

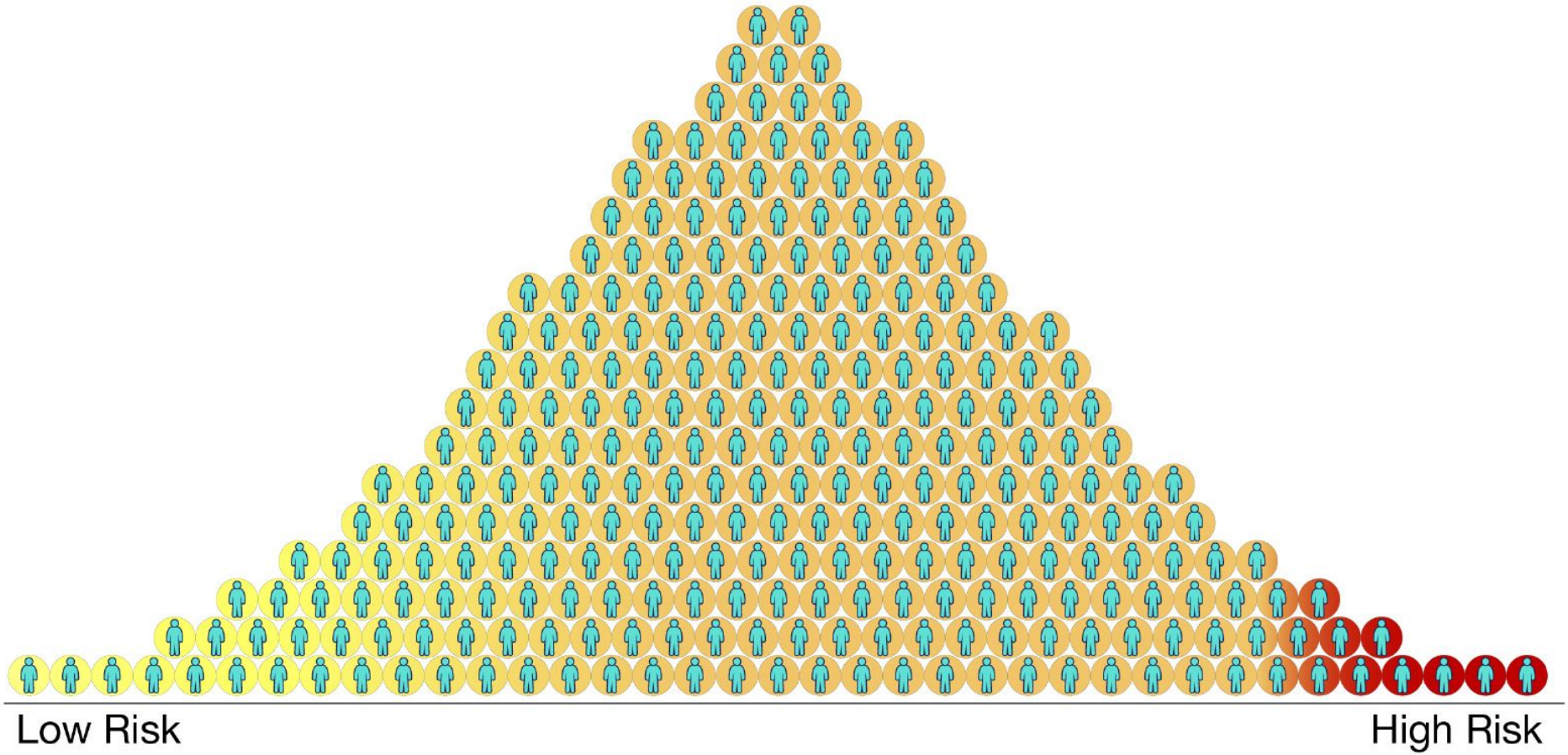


Courtesy: National Human Genome Research Institute

Polygenic Risk Score



Courtesy: National Human Genome Research Institute

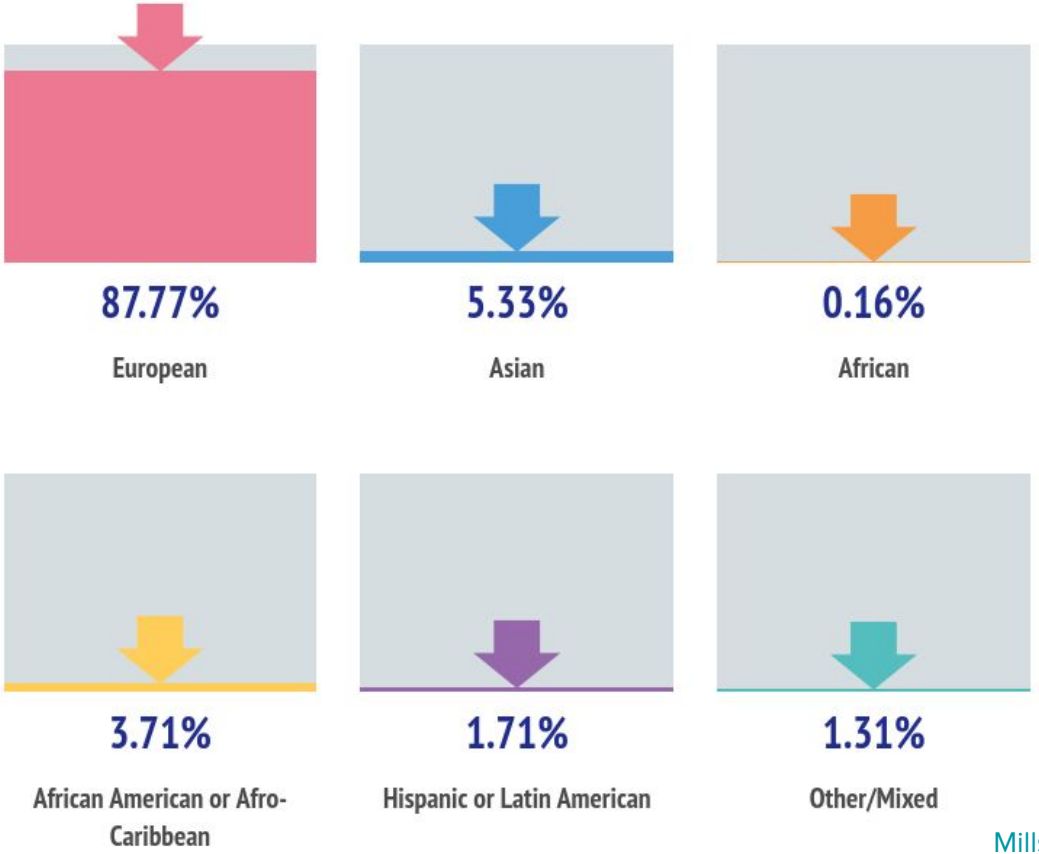


Courtesy: National Human Genome Research Institute

Genome Wide Association Studies (GWAS)

Total GWAS participants diversity

Version 1.0.0. Last check for data: 2025-09-05 00:34:26 .



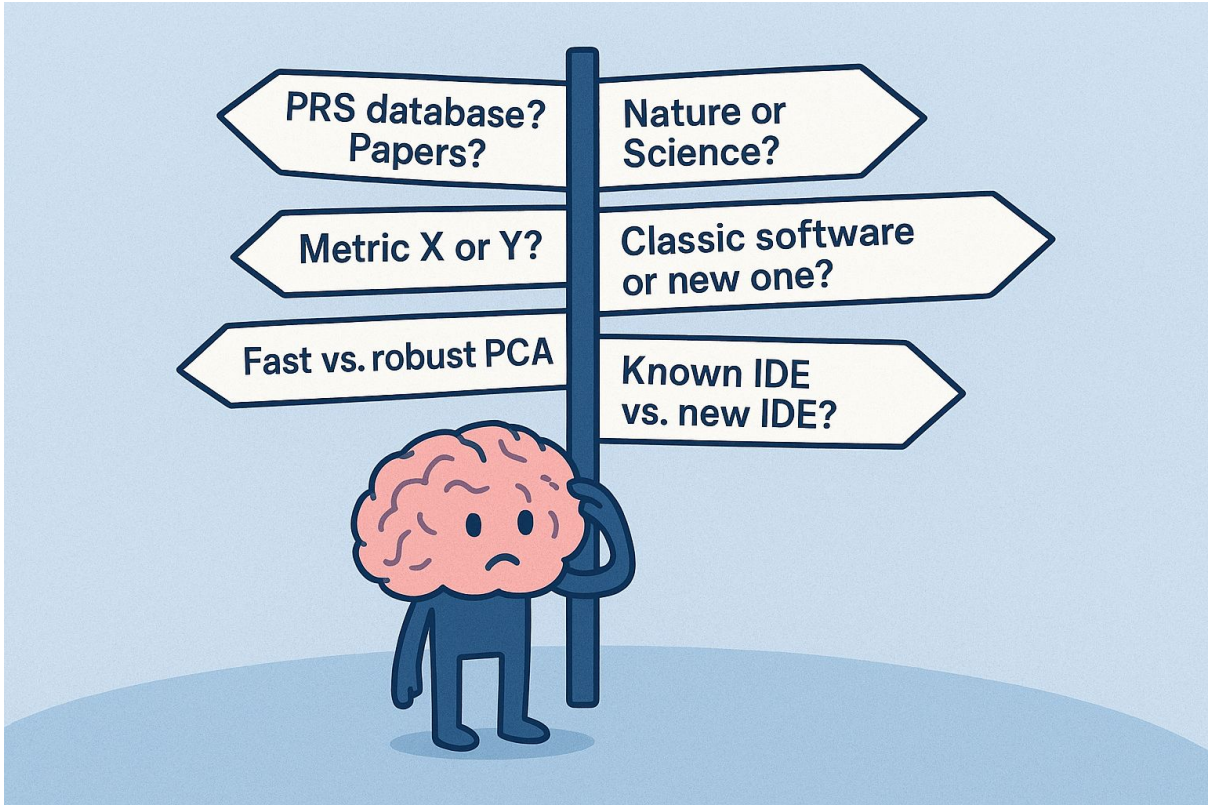
Mills, M.C and Rahal, C., (2020). gwasdiversitymonitor.com

Breast Cancer PRS Validation in Brazilians


Greek Yogurt
OIKOS

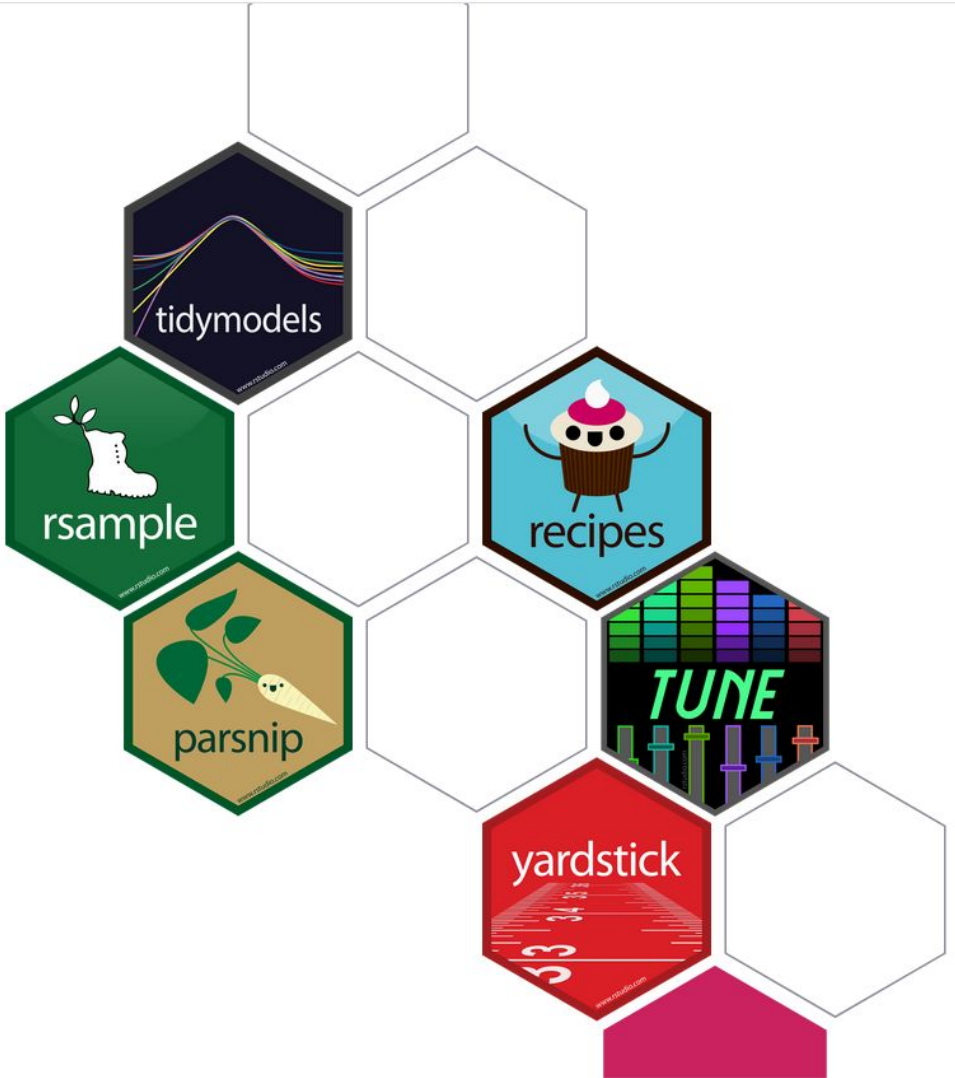

import
as

Questions, questions, questions



Can European-based PRS be used in Brazil to identify a high breast cancer risk?

Tidymodels



Tidymodels

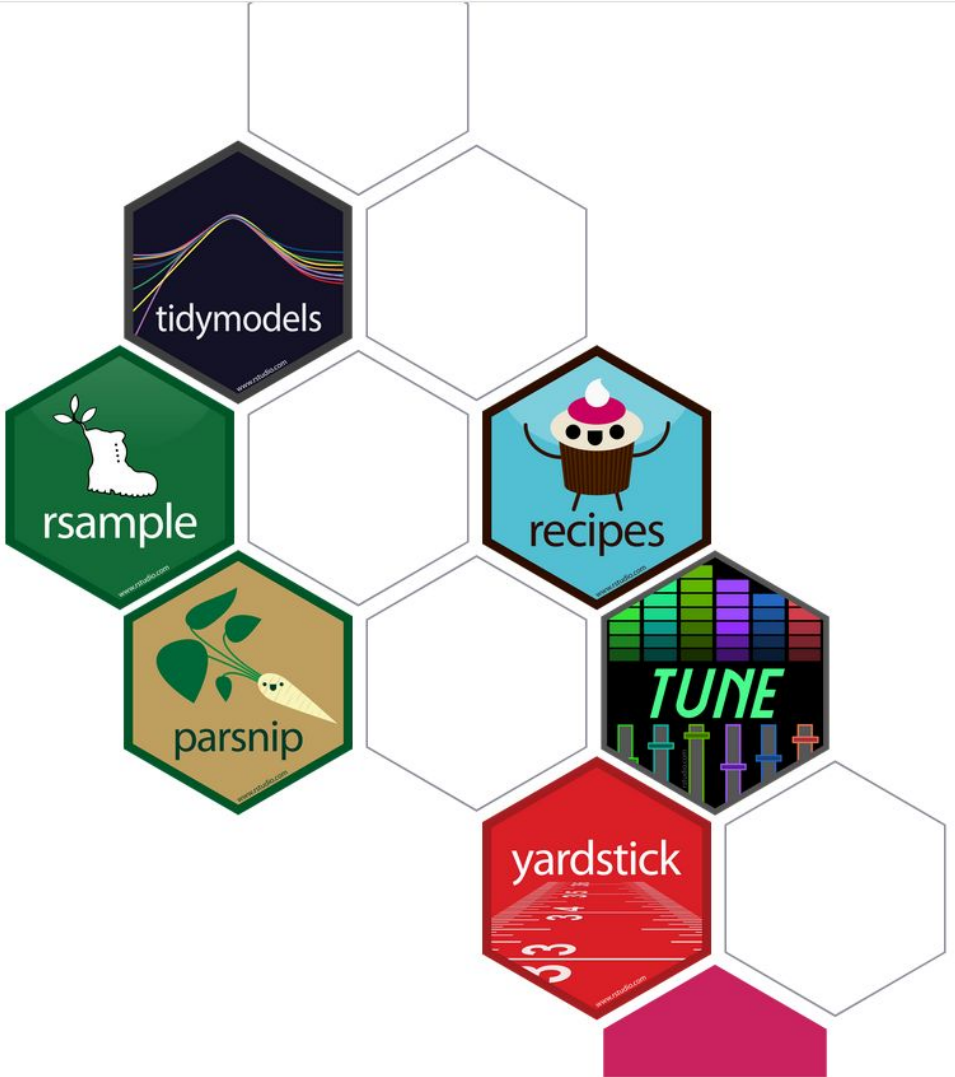
- Integratable tools

Tidymodels

- Integratable tools
- Compatible with {tidyverse}

Tidymodels

- Integratable tools
- Compatible with {tidyverse}
- Concept of reusable building blocks



Read the data

```r
breast <- readr::read_csv(here::here(
  "data",
  "output",
  "v2",
  "breast_preprocessed_no_mut_v2.csv"
))

head(breast, 3)
```

```
# A tibble: 3 × 17
  id    status sex   prs_broad prs_313 prs_3820 prs_ukbb   pc1    pc2   pc3
  <chr> <fct>  <fct>     <dbl>   <dbl>    <dbl>    <dbl> <dbl>  <dbl> <dbl>
1 1     1      F         0.501   0.488    0.505 0.000230  343. -2120. -487.
2 2     0      F         0.504   0.488    0.506 0.000234  555. -2568.  297.
3 6     1      F         0.508   0.488    0.502 0.00409  5502. -1035. -602.
# ℹ 7 more variables: pc4 <dbl>, pc5 <dbl>, pc6 <dbl>, pc7 <dbl>, pc8 <dbl>,
#   pc9 <dbl>, pc10 <dbl>
```

Split the sample

```
set.seed(28)

split <- initial_split(breast, strata = status, prop = 0.75)

train <- training(split)
test <- testing(split)
```

Normalize PRSs

```r
control_stats_train <- get_control_stats(train)

train_ctrl_stats <- train |>
  # create norm prss
  dplyr::mutate(
    prs_3820_norm = (prs_3820 - control_stats_train$prs_3820_mean) /
      control_stats_train$prs_3820_sd,
    prs_broad_norm = (prs_broad - control_stats_train$prs_broad_mean) /
      control_stats_train$prs_broad_sd,
    prs_313_norm = (prs_313 - control_stats_train$prs_313_mean) /
      control_stats_train$prs_313_sd,
    prs_ukbb_norm = (prs_ukbb - control_stats_train$prs_ukbb_mean) /
      control_stats_train$prs_ukbb_sd
  ) |>
  select(
    !all_of(c("prs_broad", "prs_313", "prs_3820", "prs_ukbb"))
  )
```

Adjust the data prior to modeling

```
rec_3820 <- recipe(
  status ~ .,
  data = train_ctrl_stats
) |>
  step_rm(
    id,
    sex,
    prs_313_norm,
    prs_broad_norm,
    prs_ukbb_norm
  ) |>
  step_normalize(starts_with("pc"))
```

Adjust the data prior to modeling

```
rec_3820 <- recipe(
  status ~ .,
  data = train_ctrl_stats
) |>
  step_rm(
    id,
    sex,
    prs_313_norm,
    prs_broad_norm,
    prs_ukbb_norm
  ) |>
  step_normalize(starts_with("pc"))
```

Adjust the data prior to modeling

```
rec_3820 <- recipe(
  status ~ .,
  data = train_ctrl_stats
) |>
  step_rm(
    id,
    sex,
    prs_313_norm,
    prs_broad_norm,
    prs_ukbb_norm
  ) |>
  step_normalize(starts_with("pc"))
```

Set engine of the model

```
log_reg <- logistic_reg() |>
  set_engine("glm") |>
  set_mode("classification")
```

Put everything together

```
wflow_3820 <- workflow() |>
  add_model(log_reg) |>
  add_recipe(rec_3820)
```

And fit

```
fit_3820 <- wflow_3820 |>
  fit(data = train_ctrl_stats)
```

Get statistics

```
1 tidy(fit_3820, exponentiate = T) |>
2   filter(grepl("prs", term)) |>
3   select(!all_of("std.error", "statistic"))
```

```
# A tibble: 1 × 3
  term          estimate  p.value
  <chr>            <dbl>    <dbl>
1 prs_3820_norm     1.44 3.15e-72
```

Update recipe to remove PCs

```
new_rec_3820 <- recipe(
  status ~ .,
  data = train_ctrl_stats
) |>
  step_rm(
    id,
    sex,
    prs_313_norm,
    prs_broad_norm,
    prs_ukbb_norm,
    starts_with("pc"))
```

```
new_wflow_rec_3820 <- wflow_3820 |>
  update_recipe(new_rec_3820)

fit_3820 <- new_wflow_rec_3820 |>
  fit(data = train_ctrl_stats)
```

Get metrics

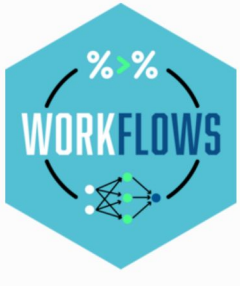
```
 1 # Predict probabilities
 2 pred_prob_3820 <- predict(new_fit_3820, new_data = test_ctrl_stats,
 3   type = "prob")
 4
 5 # Add to testing data
 6 results_3820 <- test_ctrl_stats |>
 7   bind_cols(pred_prob_3820)
 8
 9 # Obtain ROC AUC
10 roc_auc(results_3820, truth = status, .pred_0)
```

```
# A tibble: 1 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.593
```


rsample


recipes


parsnip


%>%
WORKFLOWS


broom


yardstick

Repeat with other PRSs

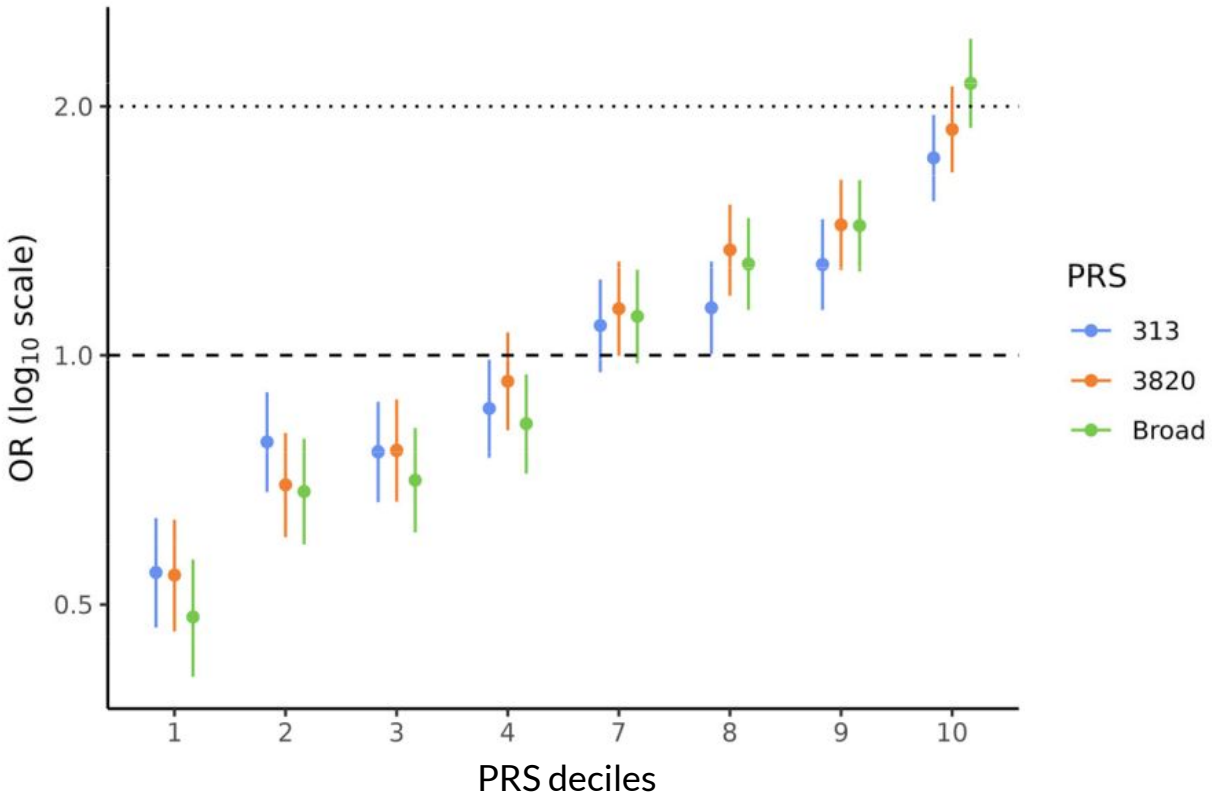
Repeat with other PRSs

Repeat by decile

Repeat with other PRSs

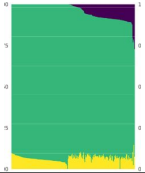
Repeat by decile

Repeat by ancestry composition


2.0
1.0
0.5
OR (log10 scale)
1
2
3
4
7
8
9
10
PRS deciles
PRS
313
3820
Broad

Per-Ancestry Results

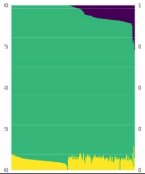
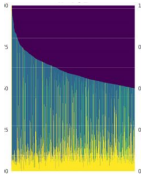
N = 214



EAS

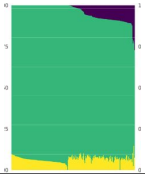
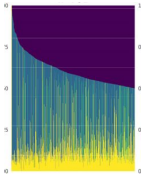
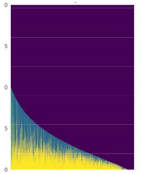
Eichemberger Rius, F. et al., 2025

Per-Ancestry Results

N = 214		**EAS**
N = 751		**Non-EUR**

Eichemberger Rius, F. et al., 2025

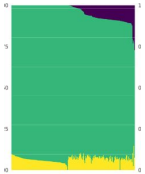
Per-Ancestry Results

N = 214		EAS
N = 751		Non-EUR
N = 13,400		EUR

Eichemberger Rius, F. et al., 2025

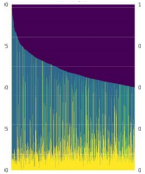
Per-Ancestry Results

N = 214



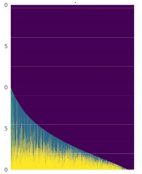
EAS

N = 751

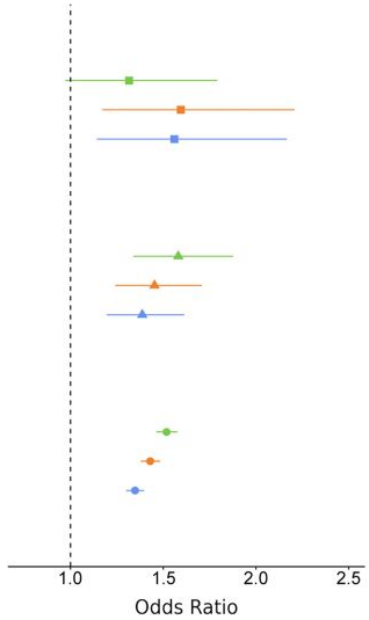


Non-EUR

N = 13,400



EUR



PRS

Broad 3820 313

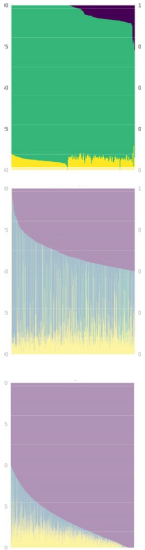
Eichemberger Rius, F. et al., 2025

Per-Ancestry Results

N = 214

N = 751

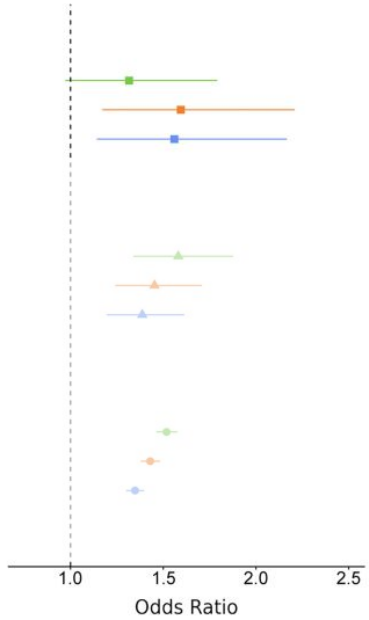
N = 13,400



EAS

Non-EUR

EUR



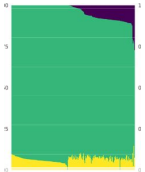
PRS

Broad 3820 313

Eichemberger Rius, F. et al., 2025

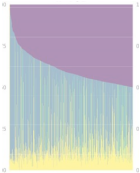
Per-Ancestry Results

N = 214



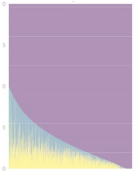
EAS

N = 751

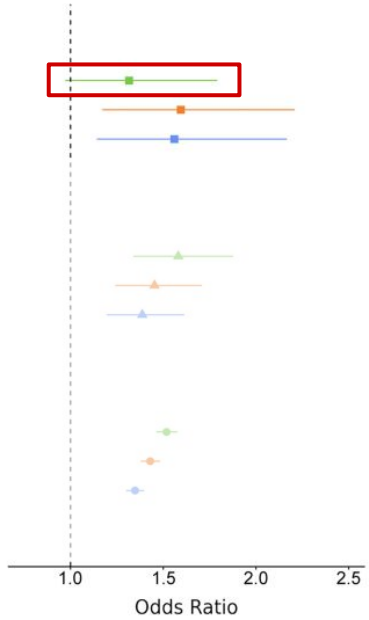


Non-EUR

N = 13,400



EUR

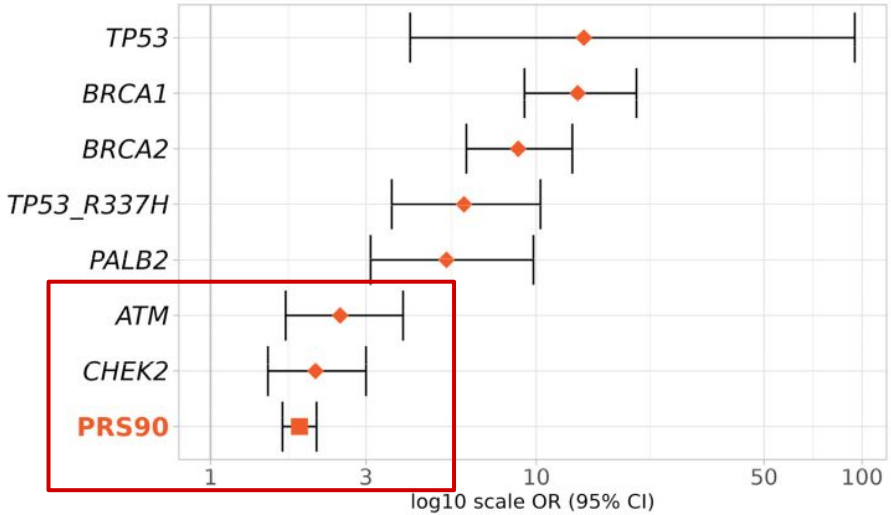


PRS

B d 3820 313

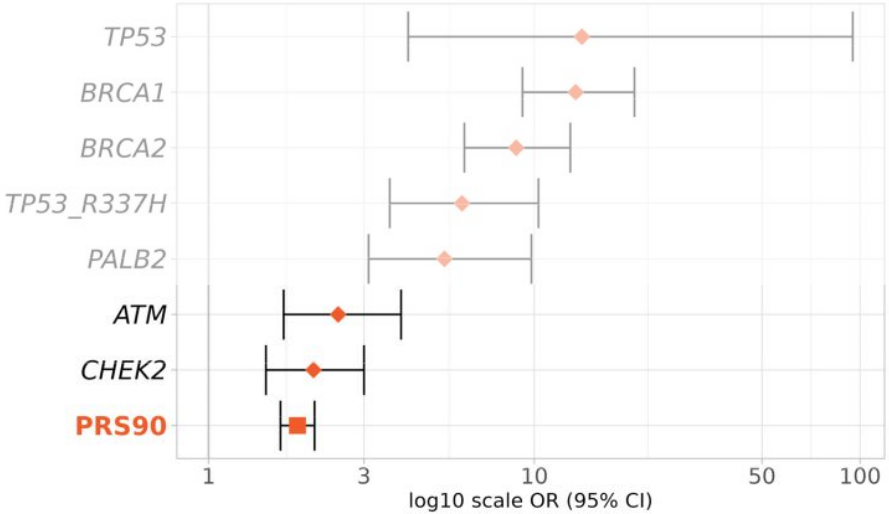
Eichemberger Rius, F. et al., 2025

PRS and Variants in Breast Cancer Genes



Eichemberger Rius, F. et al., 2025

PRS and Variants in Breast Cancer Genes



Eichemberger Rius, F. et al., 2025

diagnostics

MDPI

Article

A Breast Cancer Polygenic Risk Score Validation in 15,490 Brazilians Using Exome Sequencing

Flávia Eichemberger Rius [1,2], Rodrigo Santa Cruz Guindalini [3], Danilo Viana [1], Júlia Salomão [1], Laila Gallo [1], Renata Freitas [1], Cláudia Bertolacini [1], Lucas Taniguti [1], Danilo Imparato [1], Flávia Antunes [1], Gabriel Sousa [1], Renan Achjian [1], Eric Fukuyama [1], Cleandra Gregório [1], Iuri Ventura [1], Juliana Gomes [1], Nathália Taniguti [1], Simone Maistro [2], José Eduardo Krieger [4], Yonglan Zheng [5], Dezheng Huo [6], Olufunmilayo I. Olopade [5], Maria Aparecida Azevedo Koike Folgueira [2] and David Schlesinger [1,*]

Gold Standard Hereditary Breast and Ovarian Cancer Panel

Breast Cancer PRS

Sequencing

APC	*ATM*	*BRCA1*	*BRCA2*	*CDH1*	*CHEK2*
MLH1	*MSH2*	*MSH6*	*PALB2*	*PMS2*	*PTEN*
RAD51C	*RAD51D*	*RET*	*STK11*	*TP53*	+83 genes

Mendelics

Key Takeaways

- Projects in data science are overwhelming 🛒

Key Takeaways

- Projects in data science are overwhelming 🛒

- {Tidymodels} orchestrates ML section fluidly 🎼

Key Takeaways

- Projects in data science are overwhelming 🛒
- {Tidymodels} orchestrates ML section fluidly 🎼
- Enabled PRS validation in Brazil 🧬

Thank you!

mendelics/prs_validation_posit_conf

@flaviaerius

flaviaerius.com

Acknowledgements: *Mendelics and ICESP-FMUSP team, paper collaborators, and all patients and controls who made this study possible.*

Credits: *Angelina Jolie Stock photos by Vecteezy.*