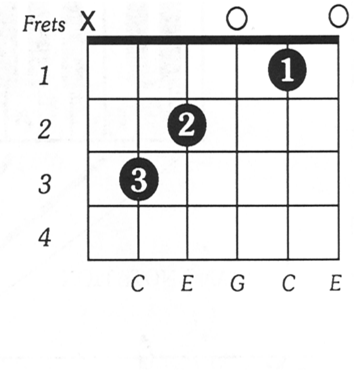
CHORD



Classifier of **HO**mologous **R**ecombination **D**eficiency

Luan Nguyen
05/04/2019

Training

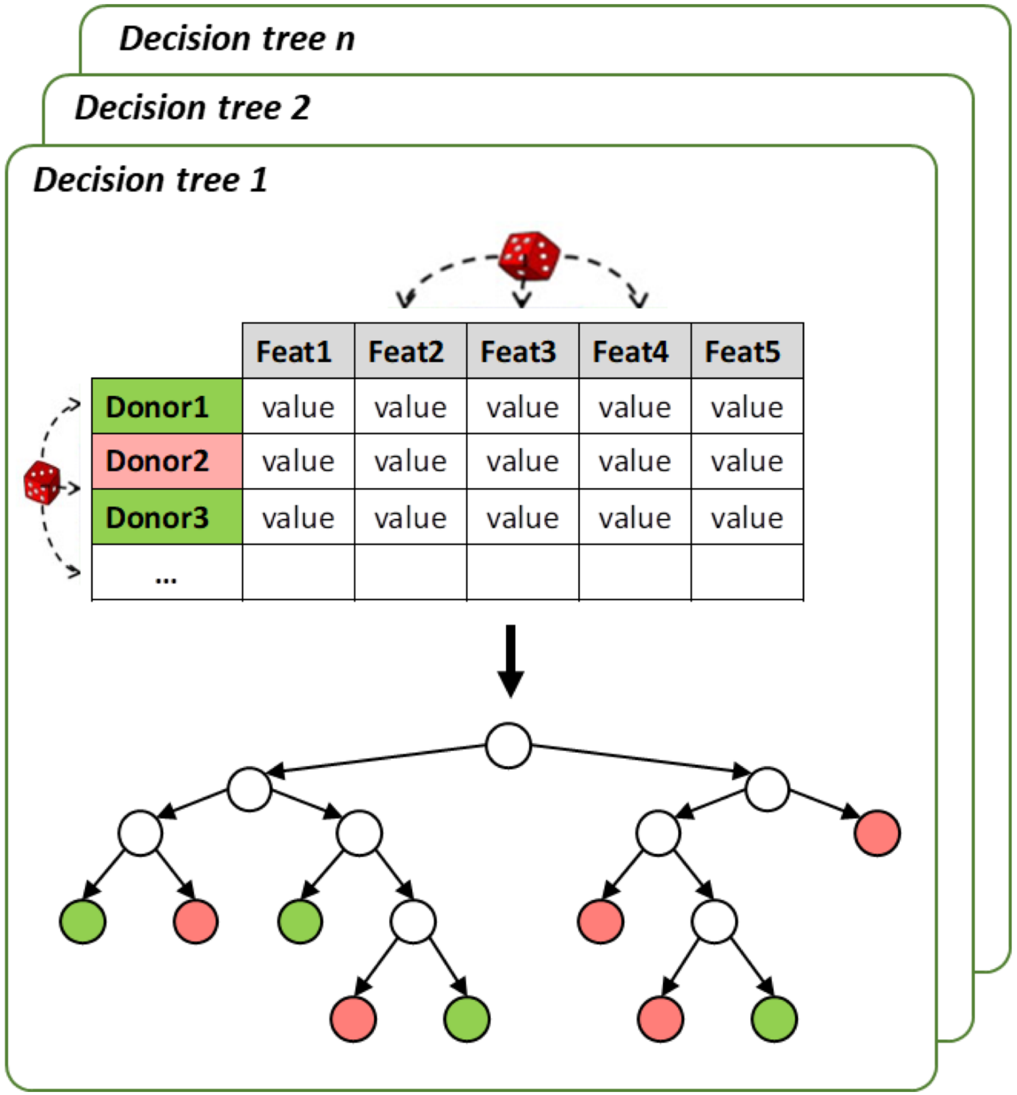
Training samples

Group	No. samples
BRCA1	25
BRCA2	65
none	1042
Sum	1132/3124

Samples originate from the Hartwig Medical Foundation (HMF)

BRCA1/2 deficient	BRCA proficient (‘none’)
For BRCA1 *or* BRCA2: • Complete loss of the gene region, *or* • LOH + pathogenic somatic mutation, *or* • LOH + pathogenic germline mutation	For BRCA1 *and* BRCA2: • No complete loss of the gene region, *and* • No LOH, *and*
Mutations: • Known pathogenic in ClinVar/ENIGMA; *or* • Frameshift	Mutations: • Benign somatic mutation or lower, *or* • Germline missense mutation or lower

Random forest



Decision tree n

Decision tree 2

Decision tree 1

	Feat1	Feat2	Feat3	Feat4	Feat5
Donor1	value	value	value	value	value
Donor2	value	value	value	value	value
Donor3	value	value	value	value	value
...					

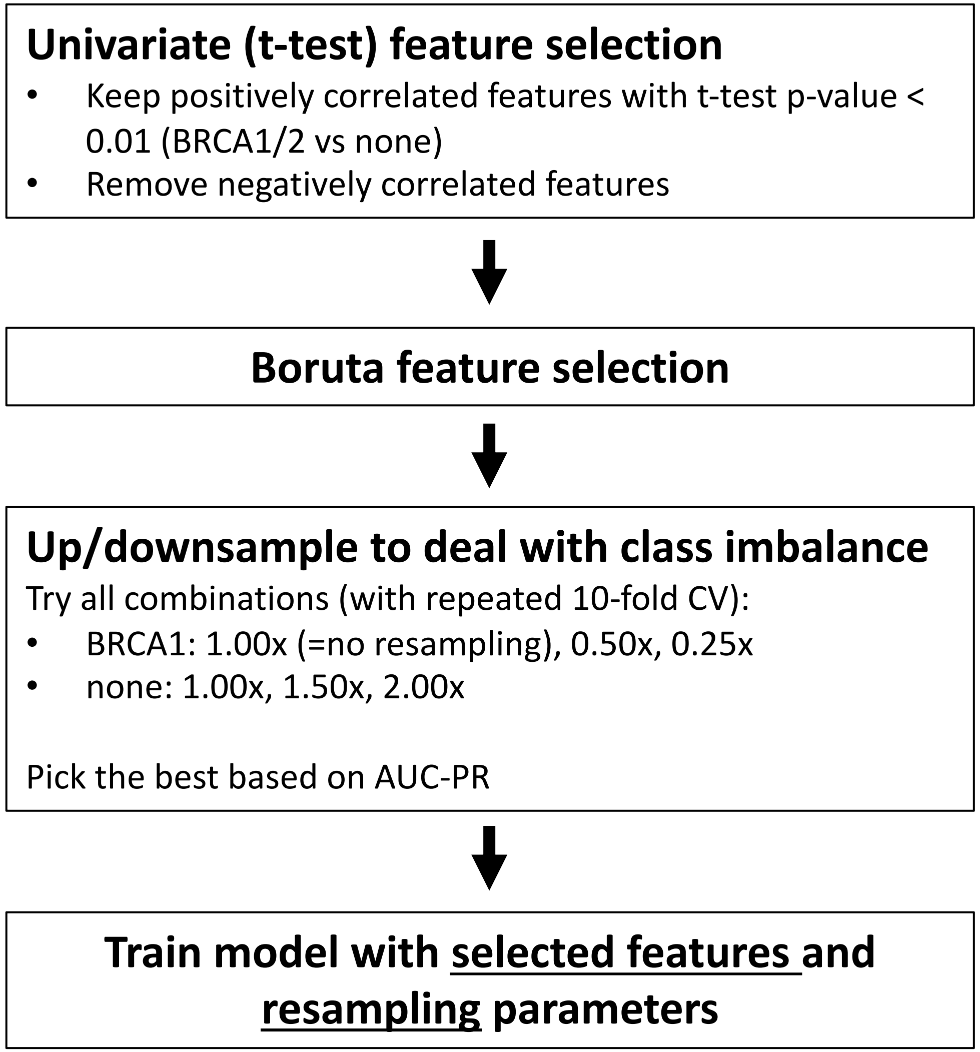
- Building one decision tree:
  - Random subset of donors
  - Random subset of features
  - Determine feature value cutoffs for branching
  - Repeat until terminal nodes are pure
- Repeat for *n* trees
- Prediction for a new sample:
  - Run feature values through each tree
  - Each tree votes BRCA1/BRCA2/none
  - $\text{Probability} = \frac{\text{Class votes}}{\text{Total votes}}$
  - Probability of HRD = $P_{\text{BRCA1 deficient}} + P_{\text{BRCA2 deficient}}$

Features

Type	Contexts	Features	No. features
SNV	Base substitution	C.A, C.G, C.T, T.A, T.C, T.G	6
Indel	• Indels within repeat regions • Indels with flanking microhomology • Other indels	• ins.rep, del.rep: (within repeats) • ins.mh, del.mh: (flanking microhomology) • ins.none, del.none: (other)	6
SV	SV type/length	DEL_0e00_1e03_bp DEL_1e03_1e04_bp DEL_1e04_1e05_bp DEL_1e05_1e06_bp DEL_1e06_1e07_bp DEL_1e07_Inf_bp … same for DUP and INV TRA (has no length)	16

Used relative contribution (per variant type) to correct
for differences in total mutational load across patients

Training procedure



Univariate (t-test) feature selection

- Keep positively correlated features with t-test p-value < 0.01 (BRCA1/2 vs none)
- Remove negatively correlated features

Boruta feature selection

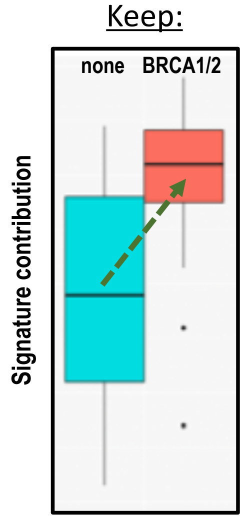
Up/downsample to deal with class imbalance

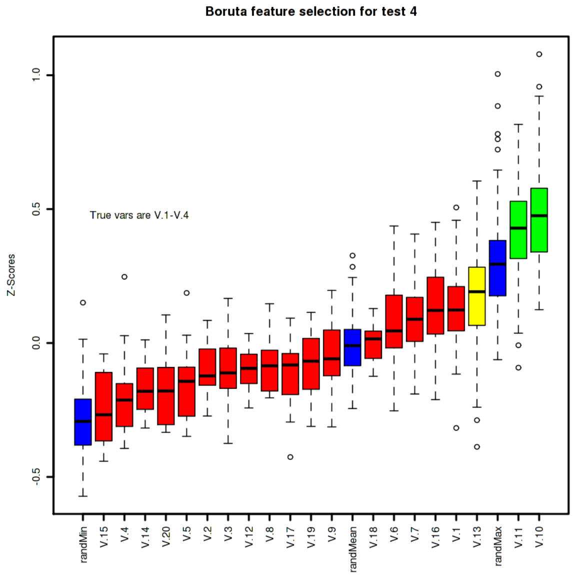
Try all combinations (with repeated 10-fold CV):

- BRCA1: 1.00x (=no resampling), 0.50x, 0.25x
- none: 1.00x, 1.50x, 2.00x

Pick the best based on AUC-PR

Train model with selected features and resampling parameters

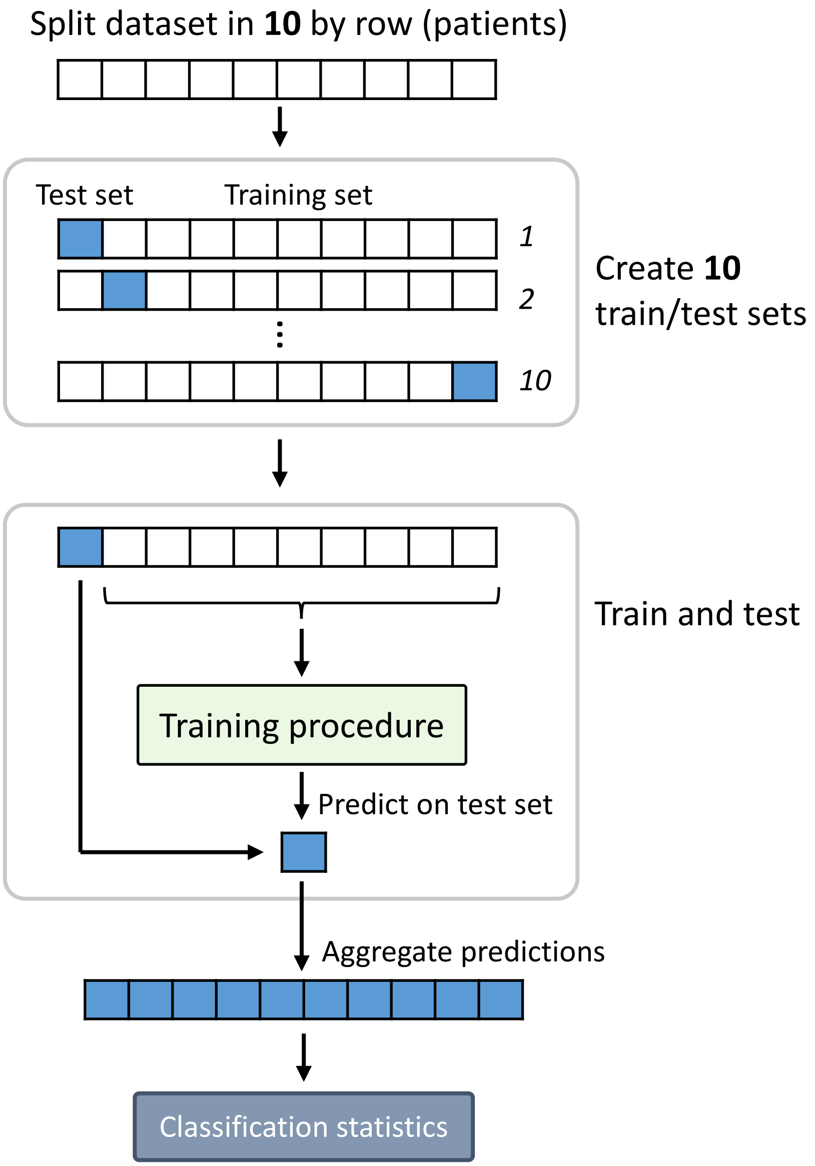




(Nested) Cross-validation

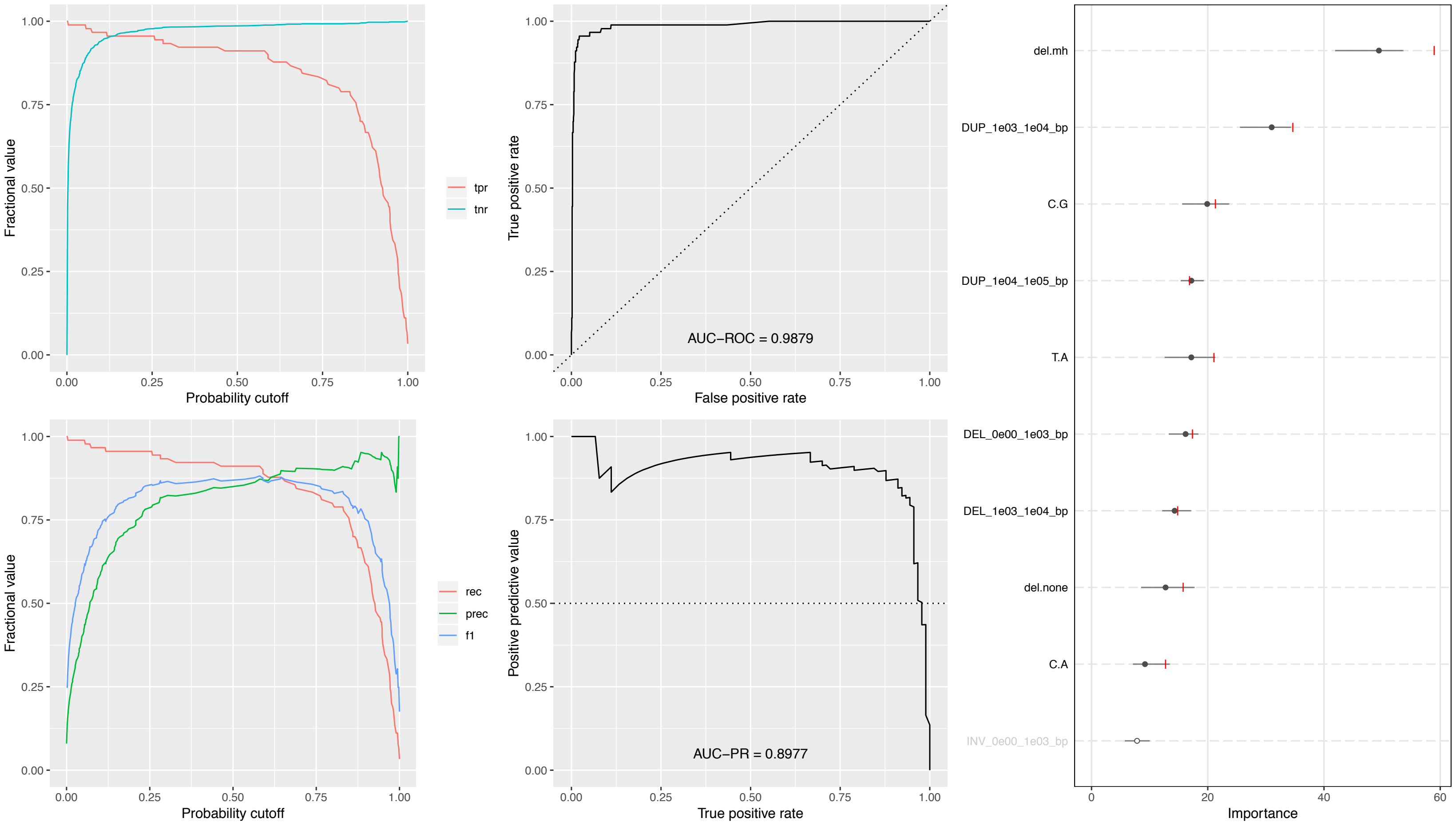
- Assessing model performance

- Predicting on 10 ‘fake’ new datasets



Performance assessed by cross-validation

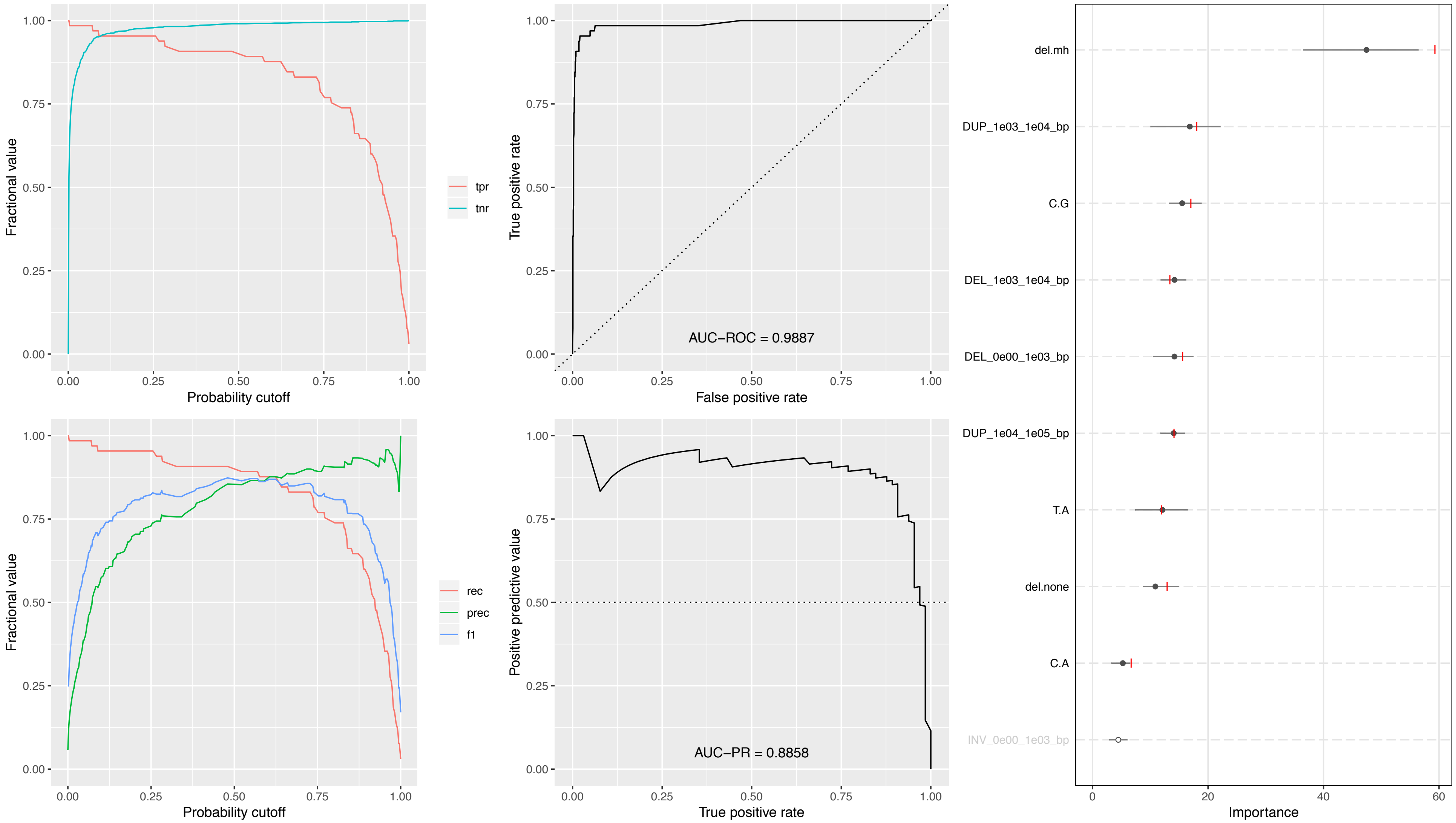
HRD prediction



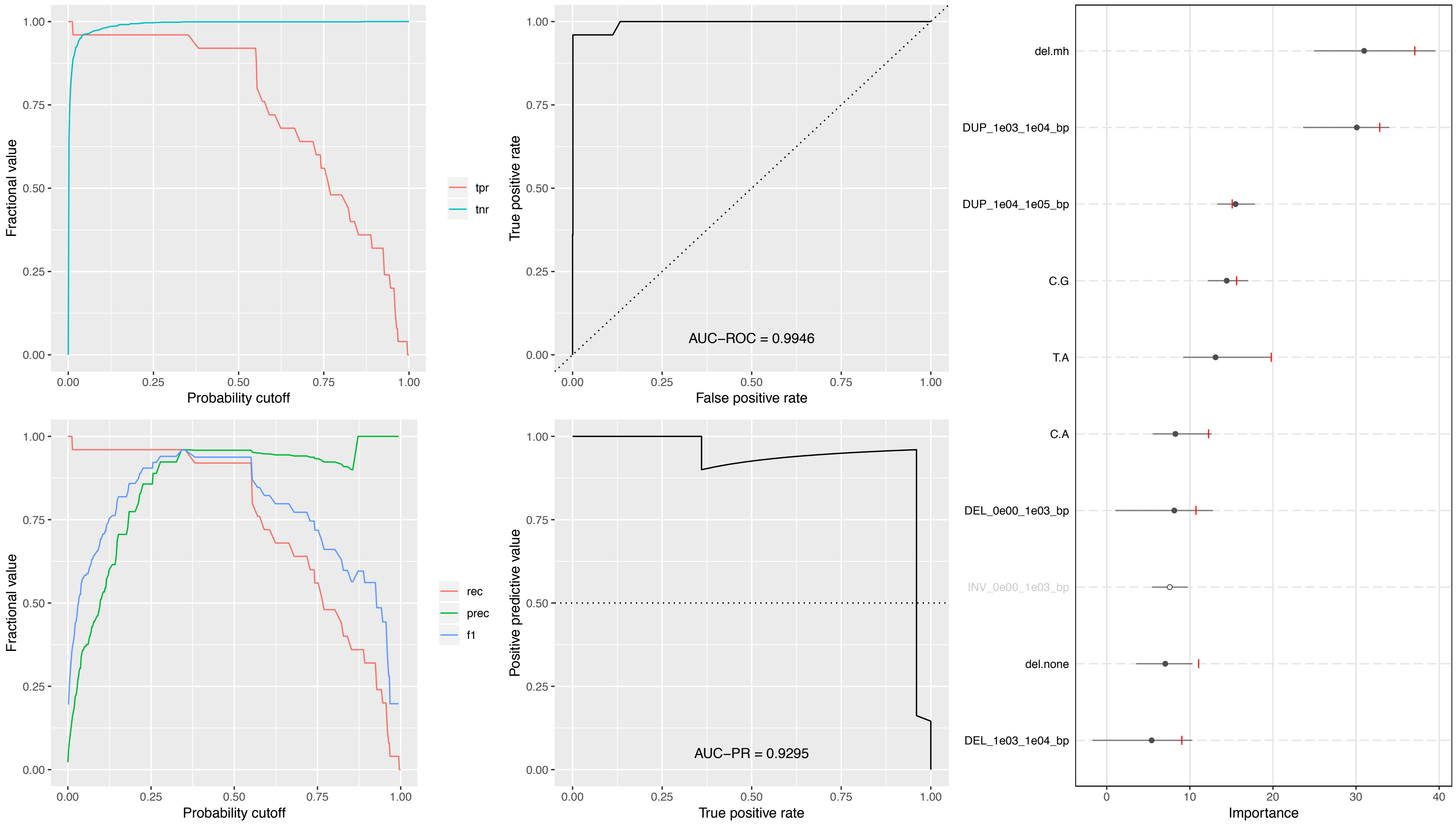
Top to bottom, left to right

- True positive/true negative rates
- ROC curve
- Feature importance
- Precision, recall, F1 curves
- Precision-recall curve

BRCA2 deficiency prediction

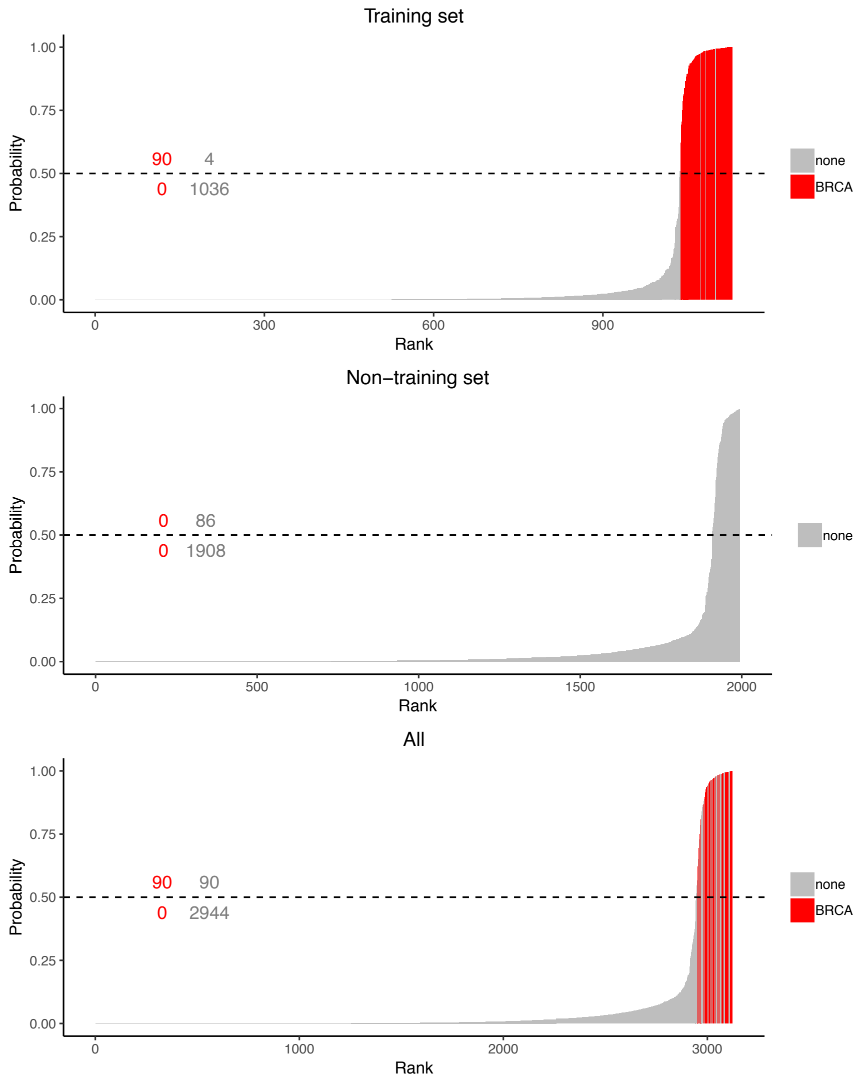


BRCA1 deficiency prediction

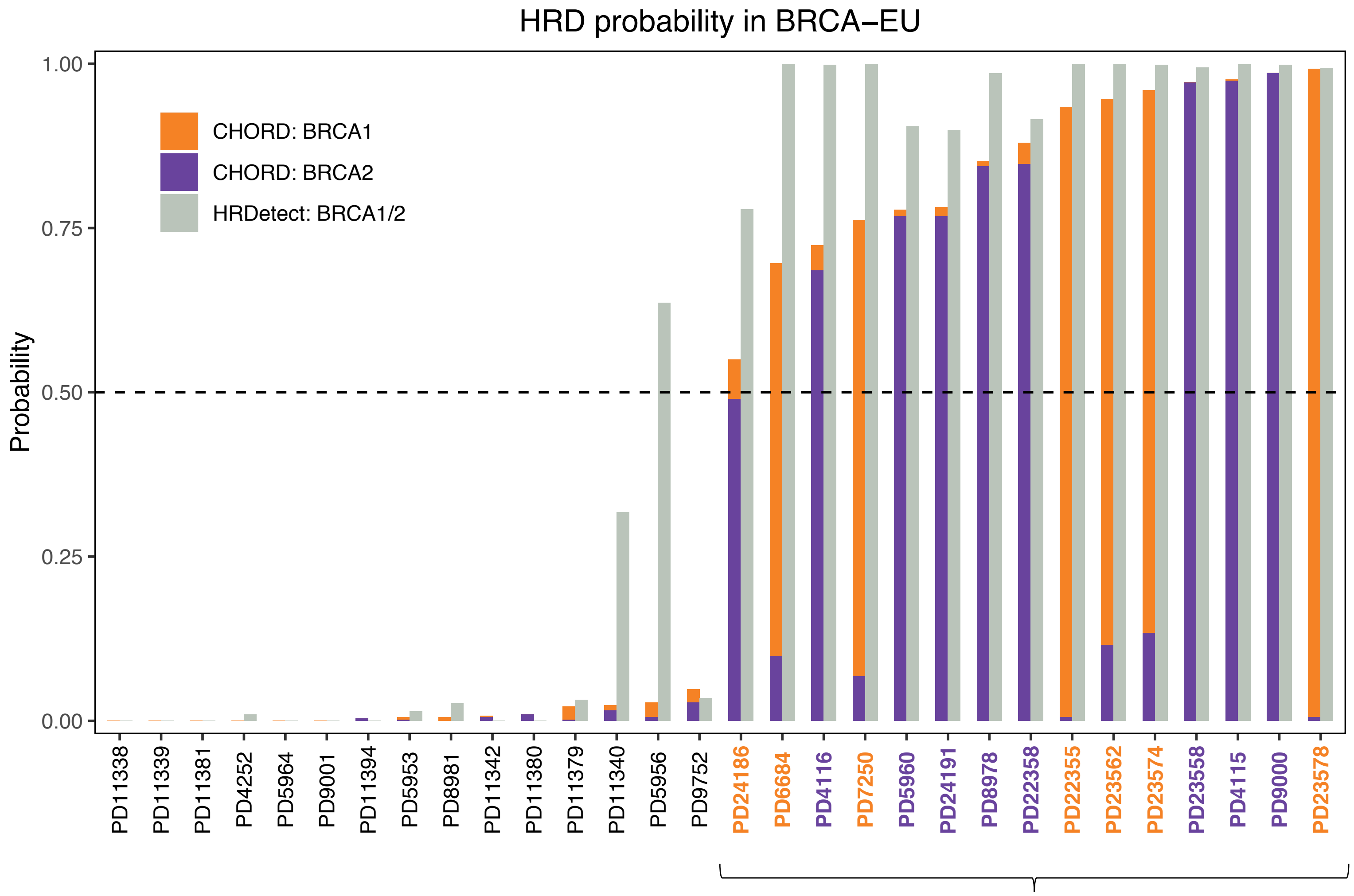


Predictions on datasets

Hartwig Medical Foundation dataset



External dataset (BRCA-EU)



BRCA1/BRCA2 annotations from HRDetect paper

- All samples annotated as BRCA1/2 deficient from HRDetect paper above cutoff
- BRCA1/2 deficiency prediction matches annotations