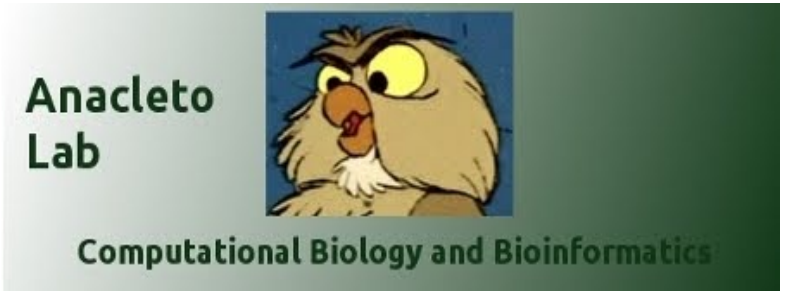
A hyper-ensemble approach for the genome-wide prediction of disease and trait-associated genetic variants

Max Schubach, Matteo Re,
Peter N Robinson, Giorgio Valentini



- Disease and trait-associated variants represent a tiny minority of all known genetic variation
- Recently ML methods have been applied to the detection and ranking of *deleterious* genetic variants in human genome
- State-of-the-art ML methods proposed in this context are not designed to deal with highly imbalanced data
- We propose *HyperSMURF*, an ensemble method designed to process highly imbalanced genomic data.

Detection of genetic variants – disease associations

Next Generation Sequencing (NGS) enables the investigation of genomic variation in coding as well in non-coding regions across the entire human genome

Application to the detection of mutations associated with Mendelian (e.g. Cystic fibrosis or Huntington disease) and *complex* (e. Alzheimer's and Parkinson's) genetic disease.

Two main problems:
1) Most of genetic variation in human genome is “physiological”: how to find “possible deleterious” variants?
2) Most studies focused on coding regions, but what about non coding regions?

Prediction of deleterious variants in non-coding genome: a challenging machine learning problem

Issues:

- How to find deleterious variants (e.g. variants associated with diseases) in the sea of physiological (neutral) genetic variation in human genome?
- A huge imbalance between deleterious (positive examples) and neutral (negative examples) variants
- Which features should be used to train learning machines for the prediction of deleterious variants?

Classical ML algorithms fail:

they are biased toward the majority class

State-of-the-art ML methods for the prediction of deleterious variants

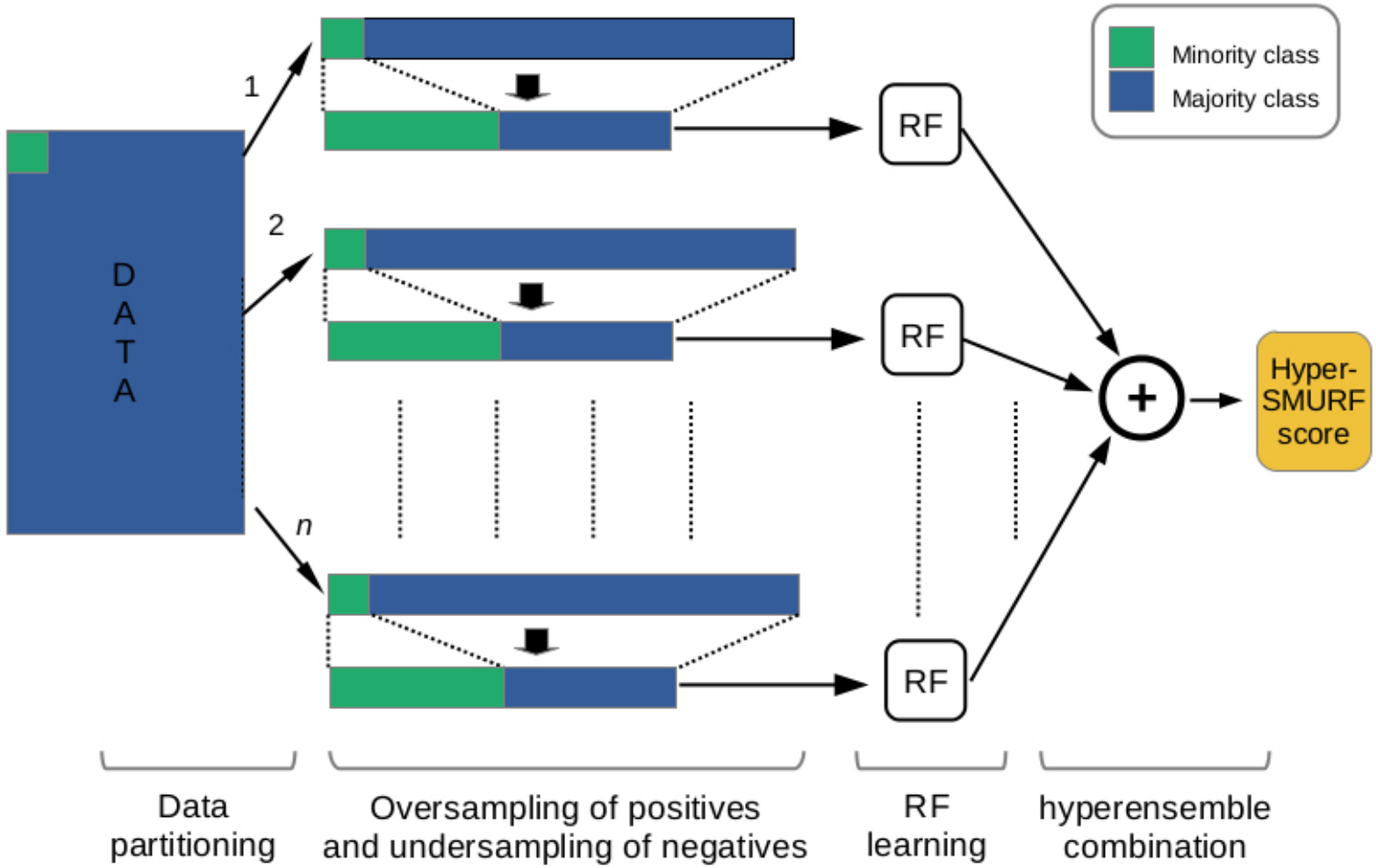
- CADD (Kircher, et al. 2014)
- GWAVA (Ritchie et al 2014)
- DeepSEA (Zhou & Troyanskaya, 2015)
- FATHMM-MKL (Shibab et al. 2015)
- Eigen (Ionita-Laza et al. 2016)

Quite surprisingly none of the above methods (apart from GWAVA) use imbalance-aware learning strategies

Our ML approach to deleterious variants detection Hyper-ensemble of Smote Undersampled Random Forests (*HyperSMURF*)

- Balancing training data through differential sampling:
  - Oversampling of the minority class
  - Partitioning and undersampling of the majority class
- Data coverage improvement and variance reduction through ensembling techniques
- Enhancing accuracy and diversity of the base learners through Hyper-ensembling

HyperSMURF:
Hyper-ensemble of SMote Undersampled Random Forests



HyperSMURF

Pseudocode of the HyperSMURF algorithm

Input:
- $\mathcal{P}$: set of positive examples (Deleterious variants)
- $\mathcal{N}$: set of negative examples (Non-deleterious variants)
- n: number of partitions
- k: number of nearest neighbors for *SMOTE* oversampling
- f: oversampling factor

begin algorithm

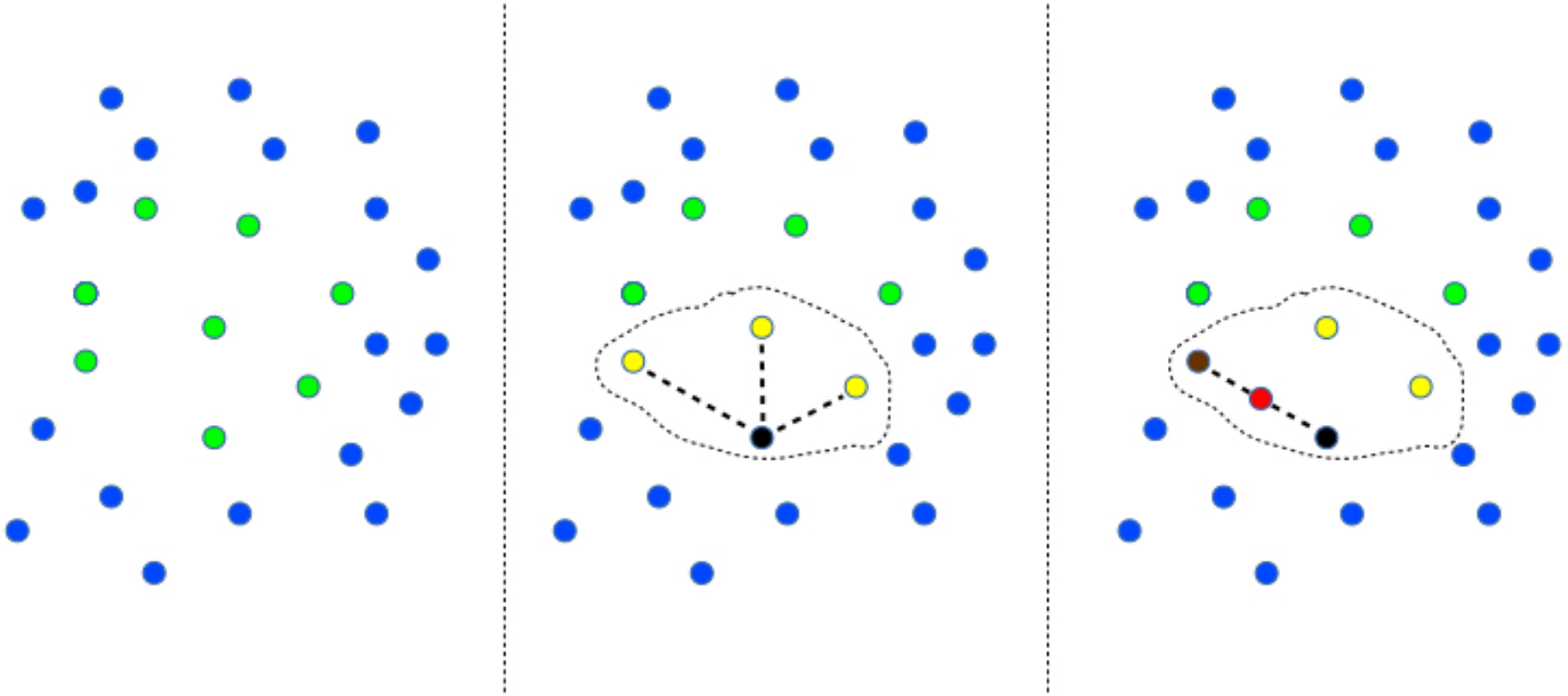
(i) Initialization and partitioning of $\mathcal{N}$:
  $n_{ex} := (f+1)|\mathcal{P}|$
  $\{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n\}$:= Do.partition $(\mathcal{N}, n)$
  $i := 1$
while $(i \leq n)$ do
  (ii) *SMOTE* oversampling:
    $\mathcal{P}_S$:= SMOTE $(\mathcal{P}, k, f)$
  (iii) Undersampling of non-deleterious variants:
    $\mathcal{N}'$:= Undersample $(\mathcal{N}_i, n_{ex})$
  (iv) Training set assembly:
    $\mathcal{T} := \mathcal{P} \cup \mathcal{P}_S \cup \mathcal{N}'$
  (v) Random Forest training:
    M_i := RF $(\mathcal{T})$
  $i := i + 1$
end while

end algorithm

Output:
$M = \{M_1, M_2, \ldots, M_n\}$: a set of RF models

Output on a test variant $\boldsymbol{x}$:
- $Hy_{score}(\boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} P(\boldsymbol{x}$ is positive $|M_i)$

SMOTE :
Synthetic Minority Oversampling Technique (*Hall et al*. 2002)



Genomic experiments

Genome-wide prediction of deleterious variants in non coding region

1) *Mendelian diseases*:
406 SNV mutations manually curated (positive examples)

14M neutral variants (negatives)

2) *Complex diseases*:
2115 regulatory GWAS hits from the GWAS catalog (National Human Genome Research Institute)

1.4M neutral variants (negatives)

Genomic attributes

1) Mendelian data: 26 genomic attributes downloaded from public data bases (UCSC, Stanford, NCBI and others):

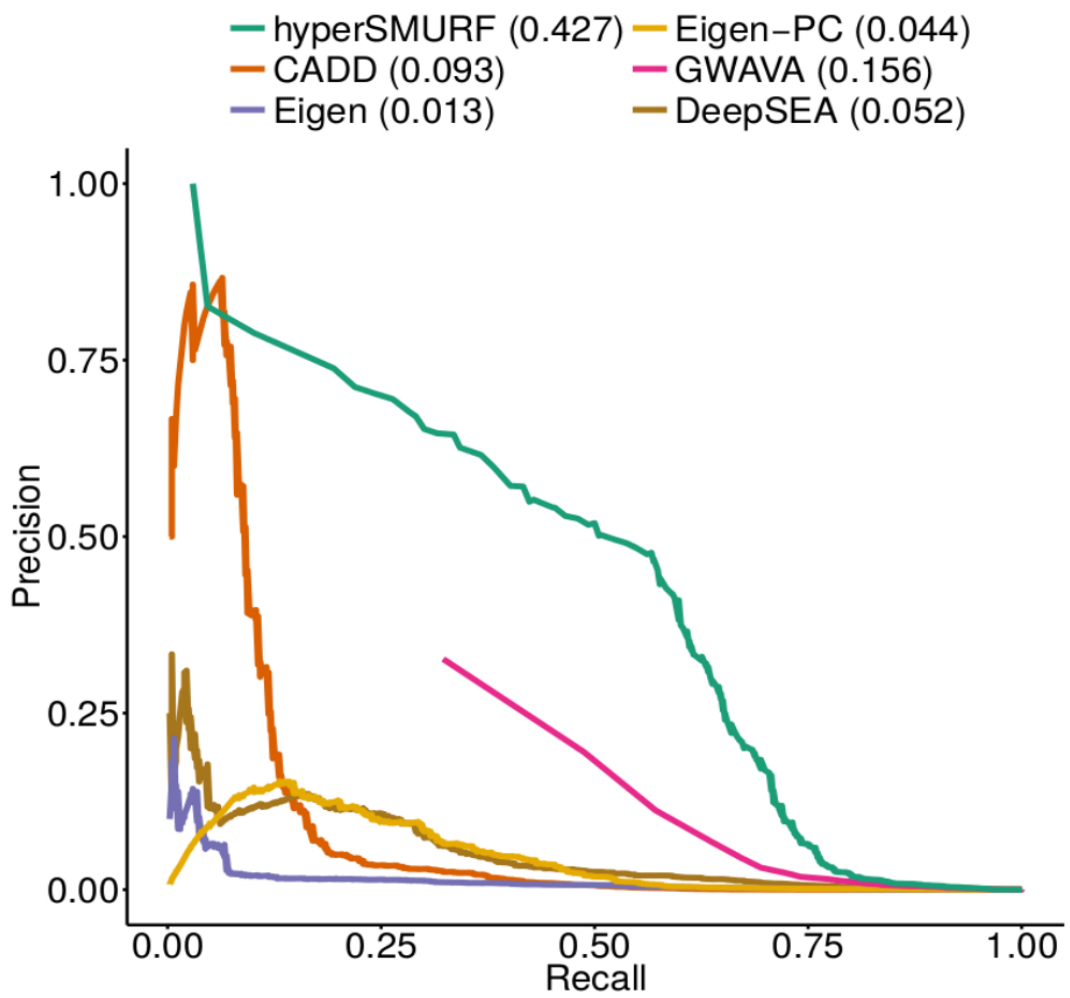
- Conservation scores
- Transcriptional features
- Regulation features
- Overalpping CNVs
- GC content
- Epigenomic features

2) GWAS data: 1842 genomic attributes directly extracted from DNA sequence through deep convolutional networks (Zhou & Troyanskaya, 2015)

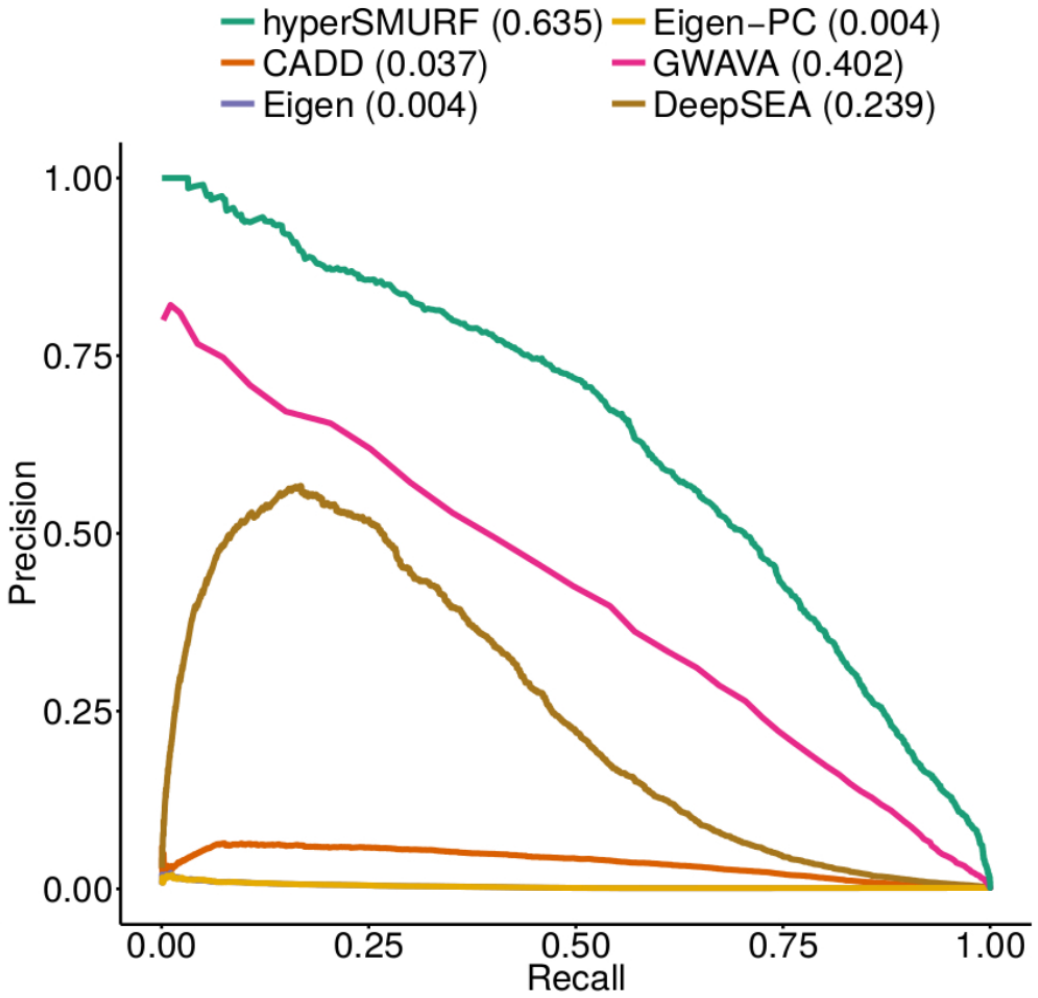
- DNAse features
- Transcription factor features
- Histone features
- Conservation scores

Comparative results with state-of-the-art methods

Mendelian diseases

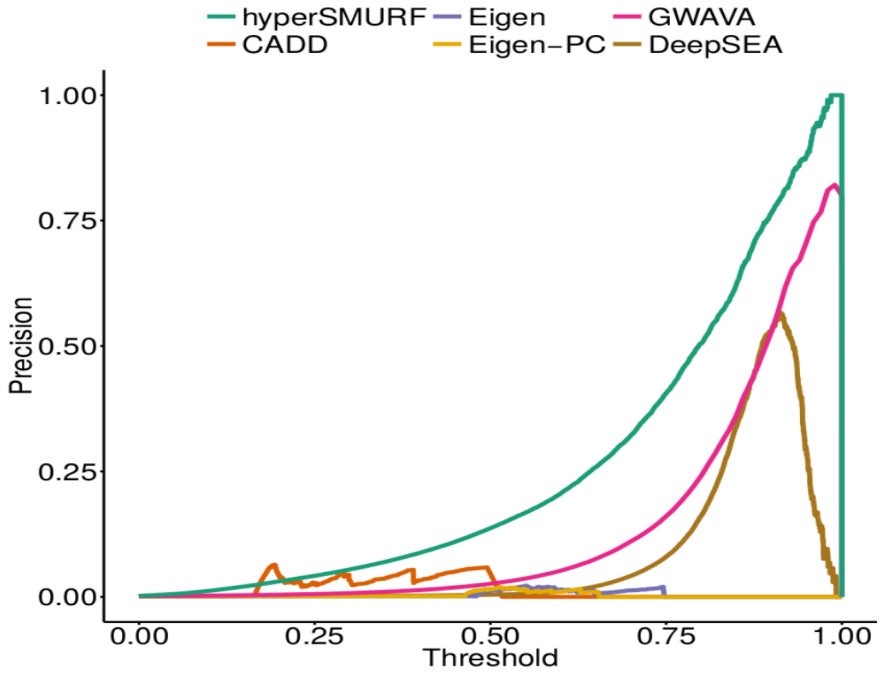


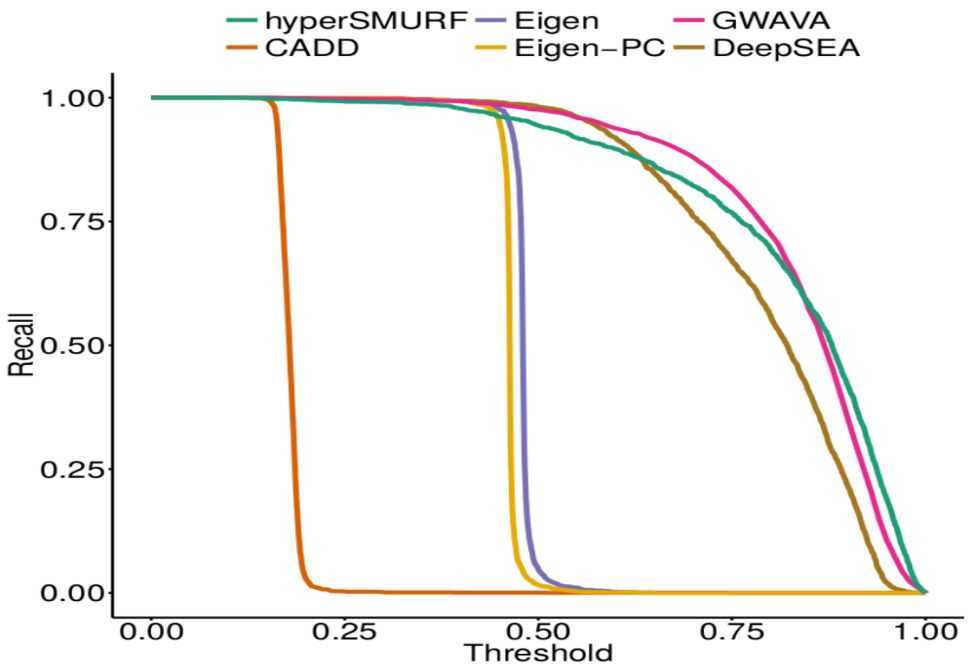
Complex diseases

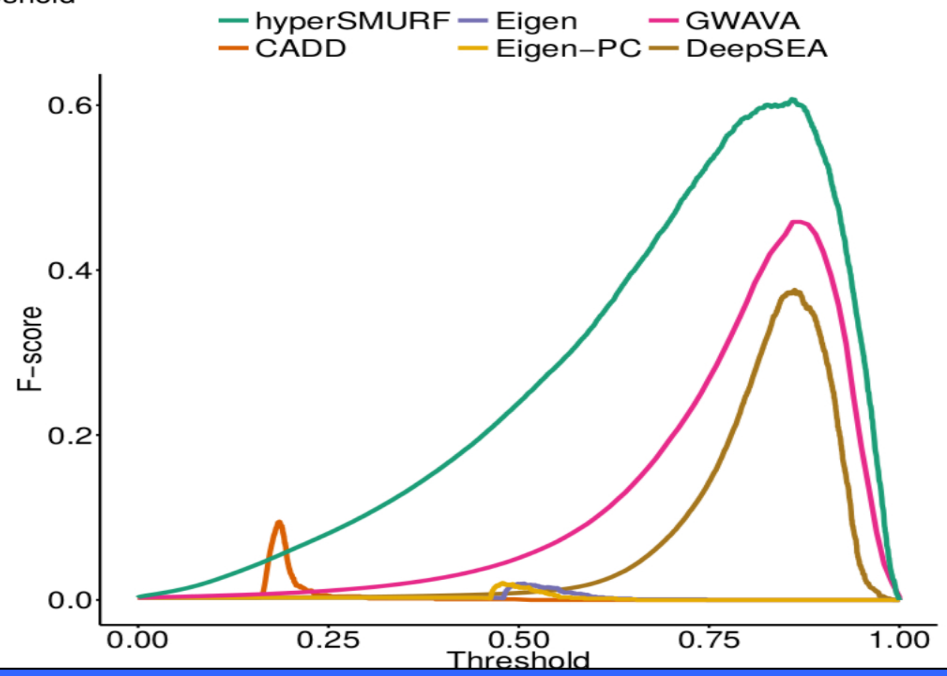


10-fold “cytoband-aware” cross-validation: precision/recall curves

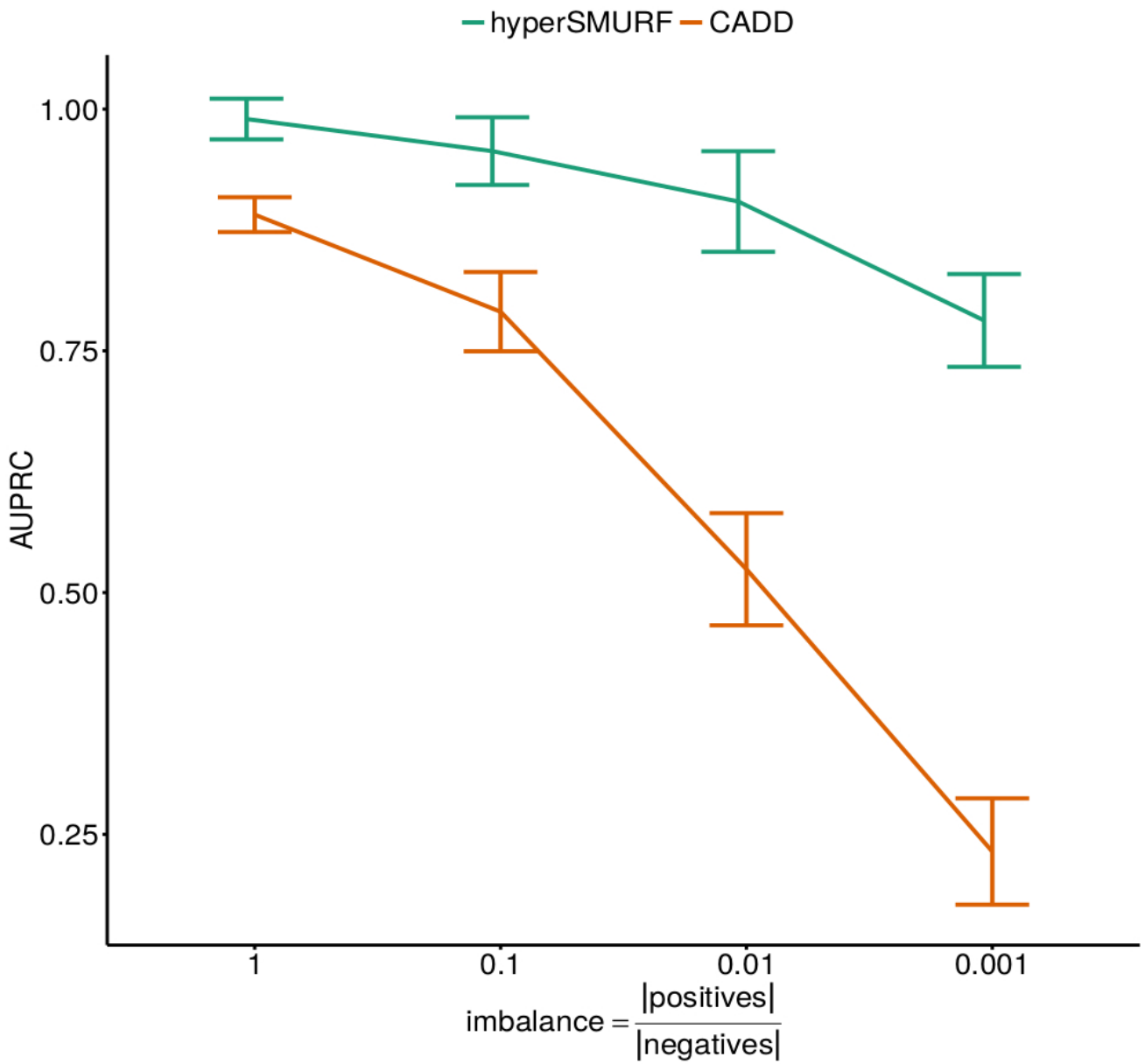
Compared precision, recall and F-score (complex diseases)







AUPRC results of HyperSMURF and CADD at different imbalance levels



Conclusions

- *HyperSMURF* is motivated by the highly imbalance that naturally arises in genome-wide studies for scoring deleterious genetic variants
- *HyperSMURF* relies on:
  a) differential sampling:
  partitioning, undersampling and oversampling techniques
  b) Ensemble methods
  c) Hyper-ensemble approach
- *HyperSMURF* software is available from:
  - https://github.com/charite/hyperSMURF (Java version)
  - https://cran.r-project.org/web/packages/hyperSMURF (R package)

References:

- M. Schubach, M. Re, P.N. Robinson and G. Valentini. *Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants,* Scientific Reports, Nature Publishing 7:2959, 2017

Nature Publishing, 7:2959, 2017

- D. Smedley, M. Schubach, J.O.B. Jacobsen, S. Köhler, T. Zemojtel, M. Spielmann, M. Jäger, H. Hochheiser, N.L. Washington, J.A. Mcmurry, M.A. Haendel, C.J. Mungall, S.E. Lewis, T. Groza, G. Valentini, P.N. Robinson. *A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease*. American Journal of Human Genetics 99:3 (2016 Sep 01), pp. 595-606.

Thank you for your attention!

http://anacletolab.di.unimi.it