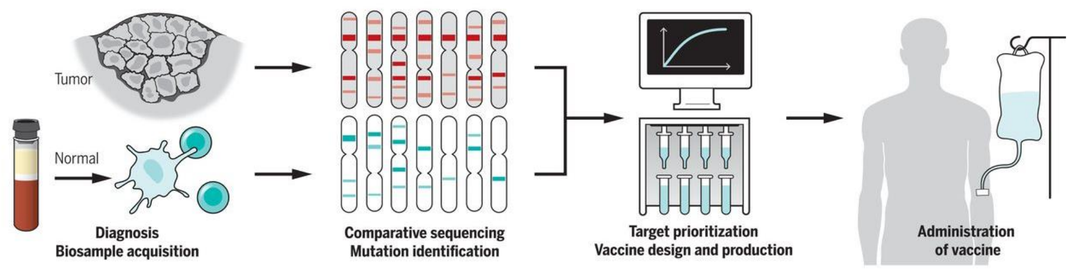
Sensitive Somatic Variant Calling Using Positive-Unlabeled Learning

Elham Sherafat, Ion Mandoiu
Department of Computer Science and Engineering, University of Connecticut, Storrs, CT

BACKGROUND

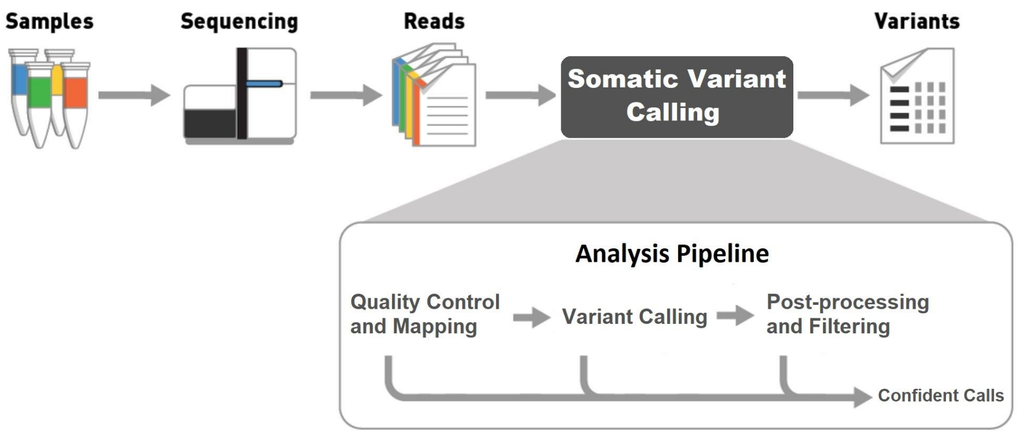
➢ Personalized cancer vaccines are emerging as one of the most promising approaches to immunotherapy of advanced cancers.



➢ This approach harnesses the power of the patient's own immune system to attack tumor cells that express specific neoepitopes generated by somatic mutations

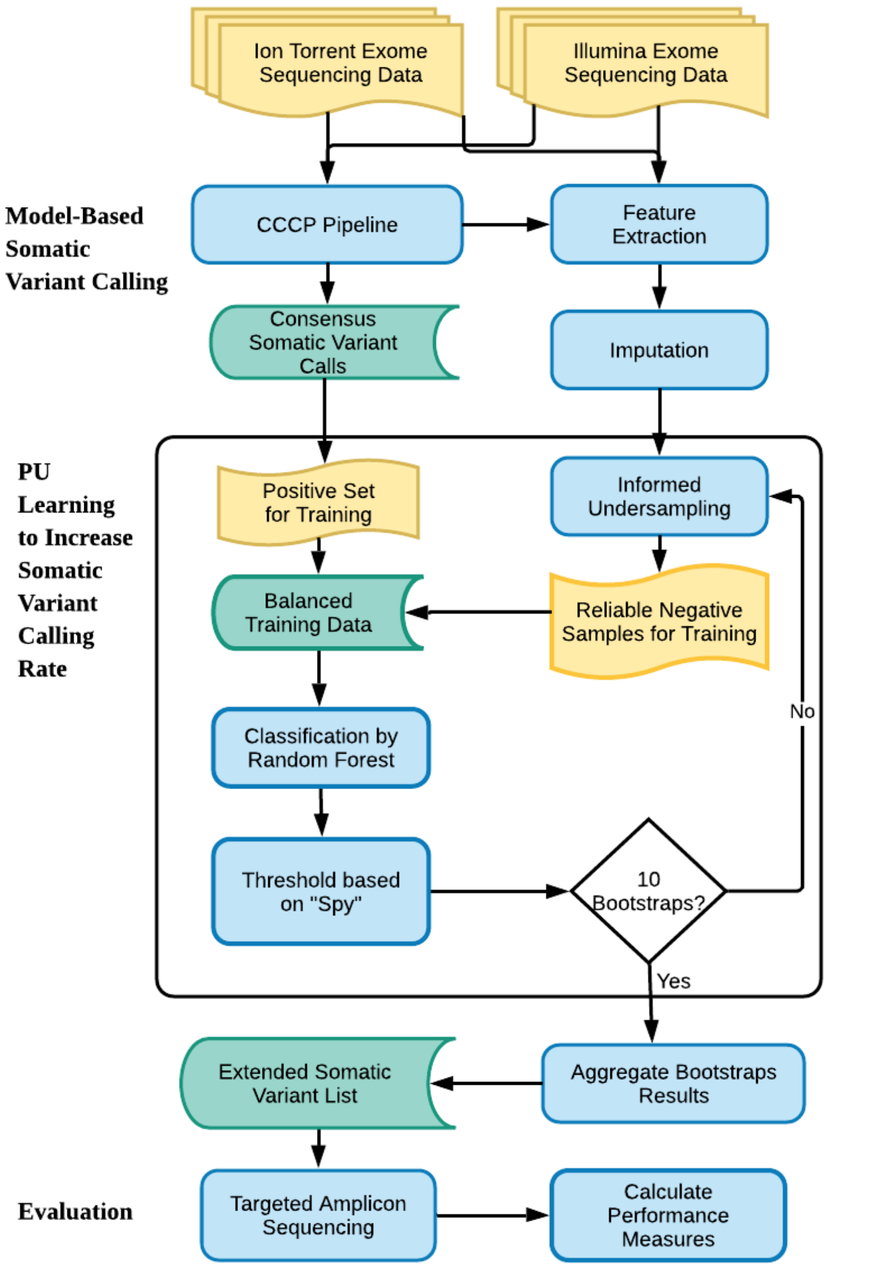
➢Somatic variant calling from matched tumor-normal next generation sequencing (NGS) data is a key step in the identification of neoepitopes that can be included in a vaccine to stimulate T-cell activation against cancer cells.

➢Although many somatic variant callers exist based on a variety of statistical models, agreement between different callers is low and accurate somatic variant calling remains challenging.



➢Key impediments to achieving consistently high accuracy with model-based methods include the large patient-to-patient variation in sample attributes such as purity, tumor heterogeneity, sequencing library preparation artifacts, sequencing errors, and errors in NGS data processing such as incorrect read alignment.

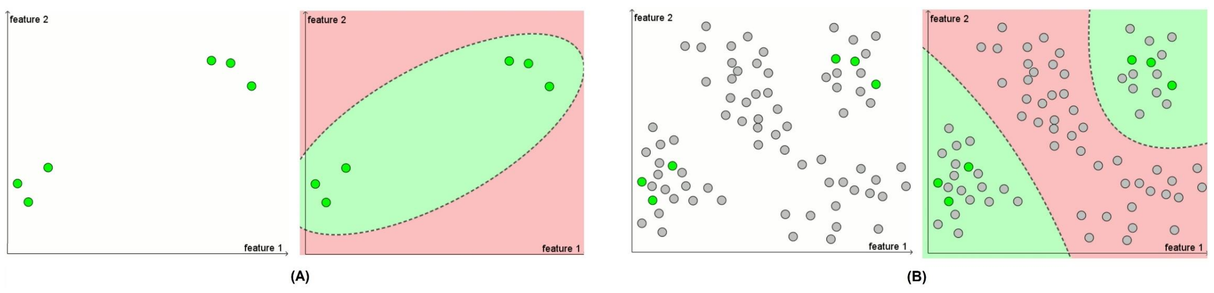
PU-CALLER WORKFLOW



OBJECTIVE

➢In this work, we use a novel machine learning method to increase sensitivity of any existing somatic variant calling pipeline while maintaining high positive predictive value (PPV).

➢To reliably handle patient-to-patient variation in sample attributes we take an unsupervised approach that learns these properties from the data itself, without a need for prior training data.



PU-CALLER METHOD

➢Two-step approach to somatic variant calling:

➢Step 1) Generate a list of confident somatic variant calls by applying an existing model-based pipeline.

➢Step 2) Extend the list of somatic variant calls using a novel *Positive-Unlabeled* (PU) learning approach.

➢In this work we use the *Consensus Caller Cross-Platform* (CCCP) pipeline in Step 1. Advantages of CCCP include:

➢Multi-technology support (can take as input Illumina and/or Ion Torrent sequencing data).

➢Incorporates *two* state-of-the-art somatic mutation callers, Strelka and SNVQ, combined using a consensus filter.

➢Previously shown to have high positive predictive value (>80%).

➢Available using easy-to-use Galaxy-based web interface at https://neo.engr.uconn.edu

➢Why PU learning?

➢Step 1 pipeline provides small number of high confidence positive calls and large number of unlabeled datapoints that fail to pass pipeline filters (unlabeled:positive ratios as large as 1000:1).

➢PU learning methods are designed for such asymmetric training data that (a) does not include negative examples, and (b) includes a very large number of unlabeled datapoints.

➢Novel aspects of PU-Caller

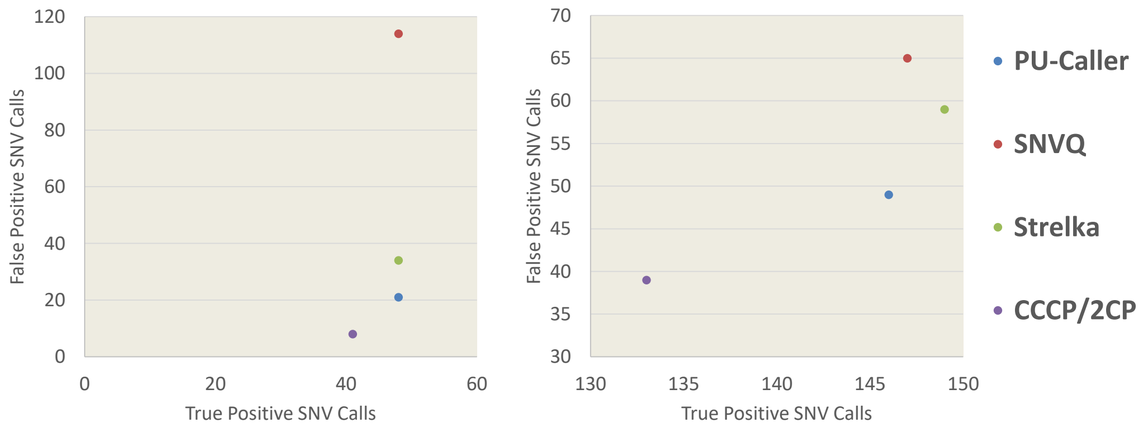
➢Dealing with data imbalance using *informed undersampling* instead of random undersampling. PU-caller partitions unlabeled data into 10 random subsets and then selects as negatives from each subset a balanced number of points that are furthest from the positive set according to the Gower distance.

➢Uses Random Forest as classifier to avoid overfitting.

➢Maintains low false positive rate by using "spy" approach in which a tenth of positive datapoints are not used for training but added to unlabeled data to determine classification threshold.

RESULTS

➢We used cancer sequencing data from two ovarian cancer patients enrolled in a Phase I clinical trial at UConn health. Matched tumor-normal exome sequencing data was generated using both Illumina HiSeq and Ion Torrent Proton sequencers. All somatic mutation predictions were validated by targeted amplicon sequencing using the AccessArray microfluidics platform.



CONCLUSIONS

➢PU-Caller yields higher sensitivity than the CCCP pipeline, with only small increase in false positive rate

➢Number of validated variants increases by 9-17% compared to CCCP, leading to additional candidate neoepitopes for vaccination

➢PU-Caller's sensitivity is similar to that of model-based callers SNVQ and Strelka but is achieved with lower false positive rate

REFERENECES

- Duitama, J., Srivastava, P., and Mandoiu, I.I. Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. BMC Genomics, 13(Suppl 2):S6 (2012)
- Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from Imbalanced Data Sets. Springer, Berlin (2018)
- Sahin, U., Tureci, O. Personalized vaccines for cancer immunotherapy, Science, 359:1355-1360 (2018)
- Saunders, C. T., Wong, W. S. W., Swamy, S., et al. Strelka: accurate somatic small-variant calling from sequenced tumornormal sample pairs. Bioinformatics, 28(14):1811 (2012)