Leveraging a comprehensive genomic data library for detecting clonal hematopoiesis in liquid biopsy

Anne Sonnenschein[1], Maysun Hasan[1], Sandra Hui[1], Bob Tell[1], Halla Nimeiri[1], Jonathan Freaney[1], Kate Sasser[1], Wei Zhu[1], and Christine Lo[1]

[1] Tempus AI, Inc., Chicago, IL

Published Abstract Number: 2324

INTRODUCTION

Clonal Hematopoiesis of Indeterminate Potential (CHIP, or CH) is a well established confounder in next-generation sequencing (NGS)-based liquid biopsy cancer diagnostics. Misclassification of CH as tumor variants can lead to false positive actionable variant detection, potentially resulting in incorrect interpretation of results and therapy selection. Moreover, CH variants may also interfere with quantitative variant monitoring leading to inaccurate assessment of treatment response. While filtering of CH is possible via matched sequencing of white blood cell and plasma DNA, emerging algorithmic approaches may enable a more resource-effective, time-sensitive approach with high precision.

METHODS

A random forest classifier was trained and validated on 1321 advanced, pan-solid tumor cancer samples (training n=660, validation n=661) sequenced using both the Tempus xF+ (liquid biopsy) and Tempus xT (solid tumor with matched buffy coat) NGS assays. Variants were labeled as CH or tumor-derived based on solid-tissue results in 39 genes that are known to be associated with CH (e.g., *DNMT3A*, *TET2*, *TP53*). The classifier was trained to classify SNV and indel variants detected via liquid biopsy as circulating-tumor or non-tumor (CH & germline) in origin. Model classifications were validated against Tempus xT.

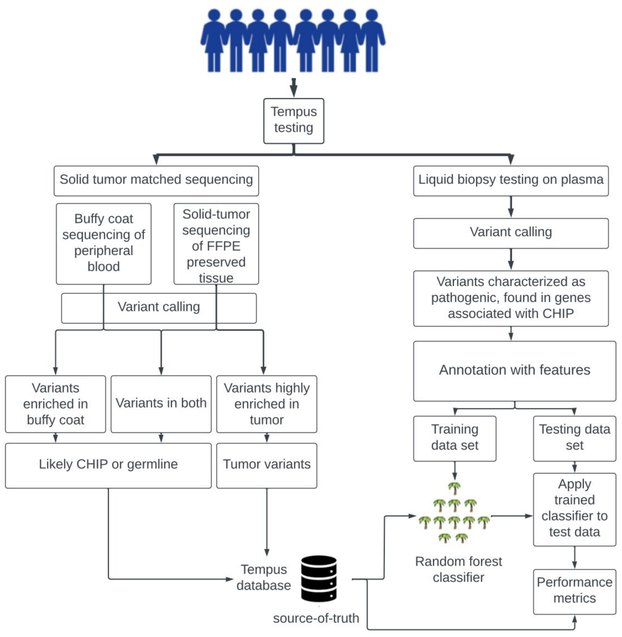


Figure 1. Data flow for liquid biopsy CH model development and testing.

SUMMARY

- A novel classifier trained on multiple orthogonal bioinformatics features can reliably distinguish CH from tumor-derived variants using only liquid biopsy data.
- An ensemble approach using multiple independent features enables high performance.
- Our classifier achieves high accuracy, including high sensitivity and high specificity.

RESULTS

Analysis of features used for classifier

Multiple features available within liquid biopsy strongly associate with CH or tumor origin, but no feature alone conclusively identifies CH variants.

Tumor fraction estimate

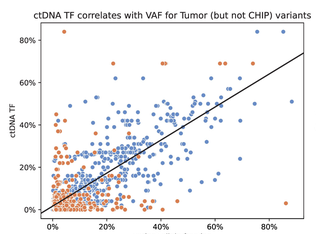


Figure 2A. The Tempus ctDNA tumor fraction estimate has a strong correlation with the variant allele fraction of variants derived from ctDNA. CH variants do not display this correlation.

DNA fragment size

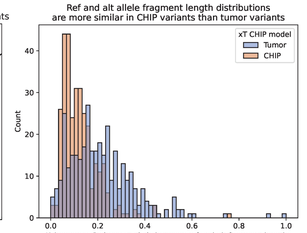


Figure 2B. DNA fragments containing circulating tumor variants have a distinct distribution relative to non-tumor. This has been previously used for tumor/CH discrimination in Marass et al., 2020.

Historical gene and variant prevalence

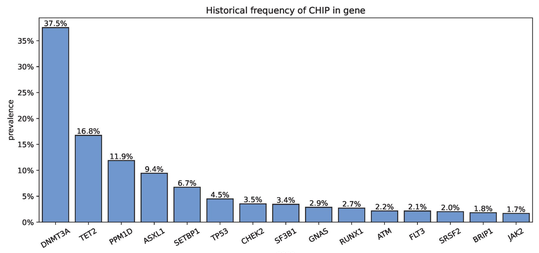


Figure 2C. CH variants in Tempus xT data by gene.

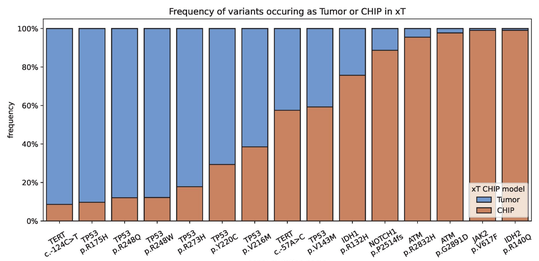
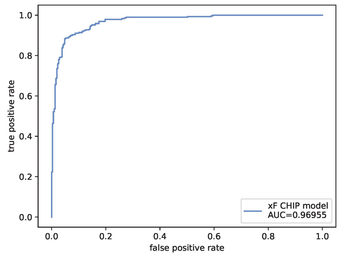


Figure 2D. Relative frequency of common CH mutations occurring as tumor derived or CH derived within Tempus xT data.

Model training, testing and performance

Metric	Training data	Test data
Sensitivity: Percentage of CH variants correctly identified	89.9%	88.3%
Precision: Percentage of predicted CH variants that are CH	95.3%	94.1%
Specificity: Percentage of tumor variants correctly identified	95.6%	94.8%



		Predicted (liquid biopsy)	
		CHIP	Tumor
Actual (solid tumor)	CHIP	TP: 257	FN: 34
	Tumor	FP: 16	TN: 293

Figure 3. Model performance on independent training and testing data. Training set was composed of 660 samples containing 680 candidate CH variants (candidates included all pathogenic variants in genes known to be associated with CH). Testing set was composed of 661 samples, containing 600 total candidate CH variants. ROC curve and confusion matrix for test data. Performance calculated at variant level. True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN).

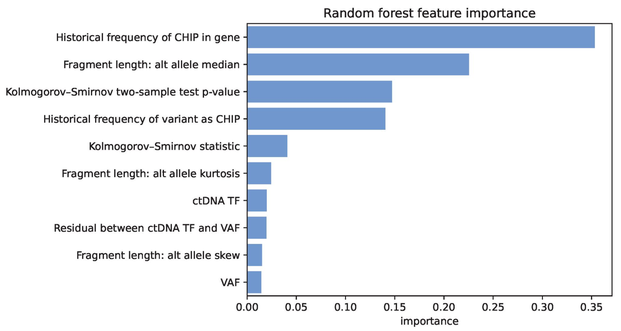


Figure 4. Feature importance in random forest model, ranked by Gini index. Gene was the dominant feature in the model, followed by statistical measures of the fragment size distribution for reads containing tumor or CHIP variants. Age is not available as a feature in this model, as it is not visible to our pipeline, which runs on de-identified data. However, age was highly correlated with CH. CH+ patients in training had a median age of 72, while CH- patients had a median age of 65 (p-value 1.1e-10 on t-test).

Characterization of identified CH variants

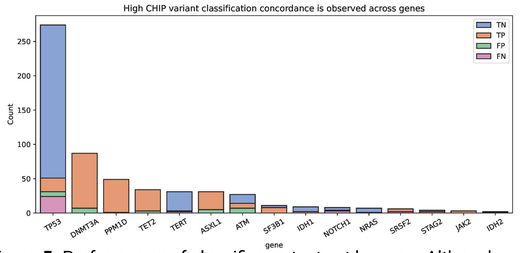


Figure 5. Performance of classifier on test set by gene. Although gene was the highest ranked feature in the model, both CH and tumor variants were seen and identified with high accuracy in many genes.

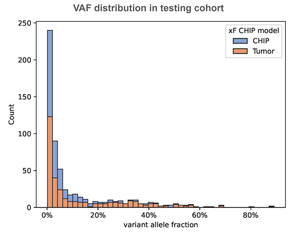


Figure 6A. Allele fractions for CH and tumor variants were similar; median VAF of 2.8% for CH and 4.4% for tumor. Although no steps were taken to exclude germline variants, their prevalence in this cohort (pathogenic variants in known CH genes) appears to be small relative to true CH.

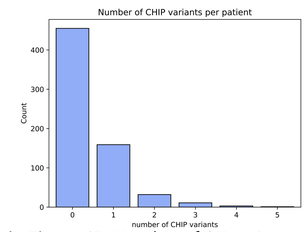


Figure 6B. Number of CH variants identified per patient. 30% of the patients used in training and testing had an identifiable CH variant. Most patients with an identified CH variant had only one variant, although a minority had multiple CH variants.

DISCUSSION

This ensemble model is highly performant at distinguishing variants derived from CH versus ctDNA, approaching accuracy previously only seen in matched or tumor-informed assays. Notably, this cohort (pathogenic variants in genes known to be associated with CH) excludes most germline variants and is naturally close to balanced between CH and tumor categories. A model for broadly distinguishing tumor and non-tumor across all genes may favor a different design. Although CH has many strongly characteristic features (association with *DNMT3A*, fragment size consistent with germline), it can present in diverse ways. Thus, accurate identification requires a multimodal approach.

ACKNOWLEDGMENTS

We thank Adam Hockenberry and Alexandria Bobe from the Tempus Scientific Communications Team for poster development support.

Correspondence: anne.sonnenschein@tempus.com