Enhancing clonal hematopoiesis variant detection in tumor-normal matched sequencing using machine learning

Anne Sonnenschein[1], Tim Baker[1], Singer Ma[1], Christine Lo[1], Bob Tell[1], Brett Mahon[1], Nirali Patel[1], Jerod Parsons[1]

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

TEMPUS

Abstract Presentation # 6344

INTRODUCTION

The Tempus xT tumor-normal matched assay, which sequences solid tumor biopsy paired with matched whole blood, has enabled the accumulation of large amounts of clonal hematopoiesis (CH) data. Although buffy coat matched sequencing is the gold standard for distinguishing tumor from non-tumor variants, accurately identifying CH variants and distinguishing them from germline or artifactual variants presents unique challenges. The buffy coat is sequenced at lower depth than the tumor, potentially impacting the accuracy of variant calling at low variant allele fractions (VAFs). For CH variants with high VAFs, which are rare but clinically relevant, distinguishing germline variants is computationally and biologically challenging. Due to immune infiltration, CH may be found in both the normal and tumor samples, and copy number variants and loss of heterozygosity (LOH) in tumor samples can substantially bias VAFs. Here, we demonstrate methods for identifying CH with high accuracy accounting for these challenges.

METHODS

A random forest model for distinguishing CH and germline was trained with variants above 20% VAF, using common variants from gnomaAD and canonical CH hotspots as a source of truth. Variants below 20% were used to train a model for distinguishing CH and sequencing artifacts, using features such as number of supporting alt reads, length of alt allele, and gene, with detection by the the higher sequencing depth xF+ liquid biopsy test as a source of truth. Combined CH calling was validated against xF+, a 523-gene liquid biopsy panel.

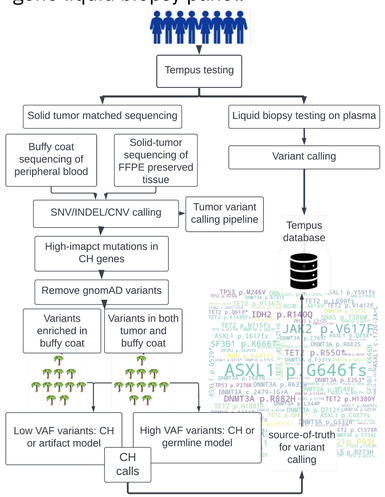

Figure 1. Model training data flow

SUMMARY

- Distinguishing high VAF CH and germline variants in tumor-normal matched sequencing is complicated by factors such as copy number variants and loss of heterozygosity
- We developed and validated an algorithm using a large dataset of germline and CH variants, demonstrating high accuracy in variant calling and germline-CH discrimination

RESULTS

Germline expectation in somatic discriminates germline and CH

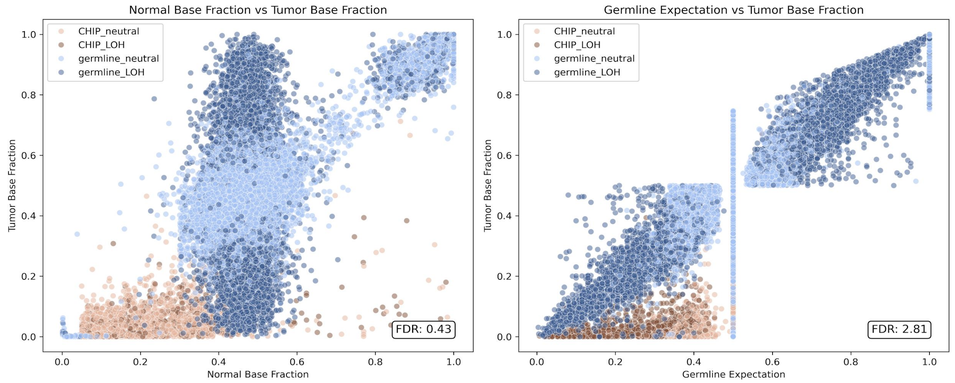

Figure 2. Germline variant VAF is consistently around 50% or 100% in buffy coat, but VAF in tumor has wide variability primarily driven by LOH (shown left), with greater variability at higher tumor purity. Incorporating copy number and tumor purity to predict expected VAF of germline variants (germline expectation) in tumor substantially distinguishes CH and germline variants (right, Fisher's discriminant ratio of 2.81) compared with ratio of germline base fraction to tumor base fraction (left, FDR 0.43).

Random forest classifier for distinguishing CH and germline using copy state

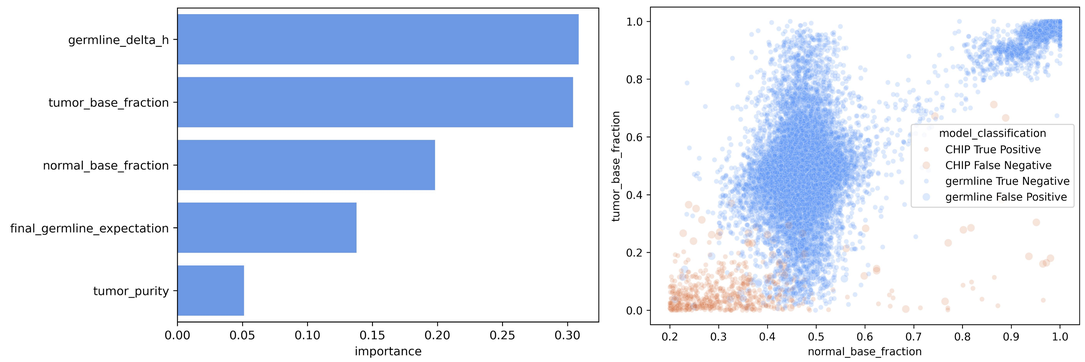

Figure 3. Random forest classifier using germline expectation feature (left) and performance shown on held out test data (right). Performance is discussed in detail in Table 1. Although the copy number informed model rescues most high VAF CH variants, a small number of variants with very high variant allele fractions in both tumor and normal are still misclassified.

Performance of CH models and validation against xF+

	CH vs germline model	CH vs artifact model	Validation against xF+
Description	**Model for distinguishing CH and germline at high (over 20%) VAF** • CH trained from 'highly canonical' CH variants (e.g. DNMT3A p.R736C, JAK2 p.V617F). • Germline trained from variants found in gnomAD at >= 10% population frequency • **ROC-AUC: 0.995**	**Model for distinguishing CH and artifacts at low (under 20%) VAF** • True CH defined as mutations detected in both xT buffy coat and somatic in xF+ assay • Artifacts defined as mutations detected in only xT buffy coat • **ROC-AUC: 0.985**	Evaluate entire CH calling pipeline, shown in **Figure 1**, tested against variant calling in xF+ using held-out validation set of 3000 unique xT/xF+ sample pairs. • True CH defined as variants that are found in xT buffy coat, are not enriched in xT tumor, and are orthogonally identified as somatic, filter-passing variants in xF+
Sensitivity (PPA)	**% of high VAF CH correctly classified as CH** • variants >= 20% VAF: 91.2% • variants >= 30% VAF: 84.3%	**% of low VAF CH correctly classified as CH** • variants < 5% VAF: 79.7% • variants >= 5% VAF: 98.3%	**% presumptive CH from xF+ correctly classified, assessed using only 'highly canonical' CH variants** • variants >= 2% VAF in xF+: 93% • variants >= 5% VAF in xF+: 95%
Precision (PPV)	**% of high VAF variants classified as CH that are not germline** • variants >= 20% VAF: 96.6% • variants >= 30% VAF: 94.4%	**% of low VAF variants classified as CH that are not artifacts** • variants <5% VAF: 85.8% • variants >= 5% VAF: 95.6%	**Percentage identified as CH in xT and classified as somatic in xF+** • variants >= 2% VAF in xT buffy coat: 90.3% • variants >= 5% VAF in xT buffy coat:: 93.5%

Table 1. Evaluation of model components and final CH calling pipeline

Characterization of CH validated against xF+

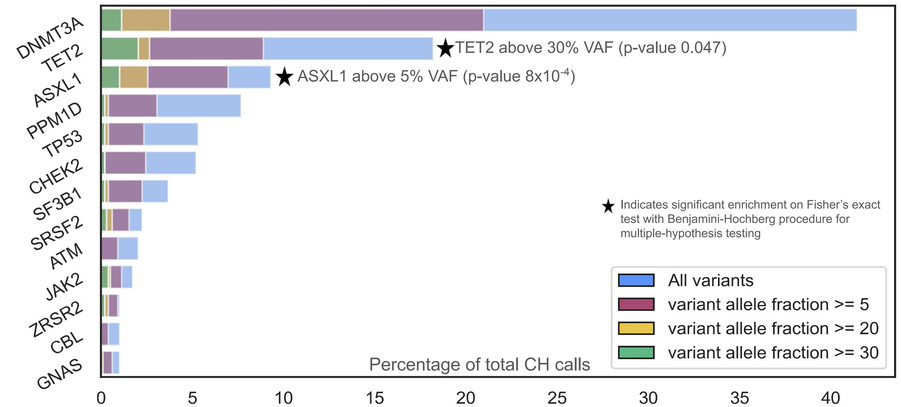

Figure 4: Variants correctly classified as CH follow the expected gene distribution. However, amongst very high VAF CH, the distribution shows differences. Very high VAF CH is rare overall, but occurs frequently in high throughput data. 3.6% of assessed samples have a CH variant over 20% VAF, while 1.7% have a variant over 30% VAF. Presence of complex clonality is moderately associated with high VAF variants; 23% of samples with 3 or more CH variants have at least one variant >= 30% VAF, compared with 4.4% of samples with only 1 variant (p-value 1.33×10^{-5}, Fisher's exact test). Age is slightly correlated with number of variants found in a sample (Spearman Correlation 0.23, P-value 1.4×10^{-6}).

ACKNOWLEDGMENTS

We thank Dana DeSantis from the Tempus Science Communications team for poster development.

Correspondence: anne.sonnenschein@tempus.com