Combining State Space Models and Statistical Decision Theory for Precision Profiling of Thalassemia

Austin Talbot[1]*, Alex V. Kotlar[2], Lavanya Rishishwar[1], Andrew Conley[1], Mengyao Zhao[1], Nachen Yang[1], Michael Liu[1], Brian Wright[1], ShiPing Zou[1], Sean Polvino[1], Yue Ke[1], Zhaohui Wang[1]

[1]Pillar Biosciences Inc., [2]RevXon. *Corresponding author talbota@pillarbiosci.com

Pillar Biosciences, Inc. 9 Strathmore Rd Natick, MA 01760
(800) 514-9307 | www.pillarbiosci.com

ABSTRACT

Introduction. Liquid biopsy and targeted NGS have emerged as economical and effective options for disease diagnosis, particularly genetic disorders. However, detecting copy number variations (CNVs) and associated disorders is a significant challenge. PiVAT® (Pillar's Variant Analysis Toolkit) with our inheritReveal™ Thalassemia RUO panel currently has capabilities to help labs detect thalassemia that has performed well previously, however, it has struggled with overconfidence in clinical normals. We developed a Bayesian model to estimate the sample CN ratios along with an annotation on thalassemia type. A state space model is used to estimate the CN ratio at each exon, these estimates are then used to profile the type of thalassemia using statistical decision theory. We can provide uncertainty quantification for both the annotations and the copy ratio estimates, along with quantitative metrics for sample quality and confidence, making it a robust tool for precision molecular testing.

Methods. We analyzed 13 normal controls and 44 samples with different forms of alpha thalassemia. To augment our dataset, we performed 4-fold evaluation, using two of the normals for evaluation per run, resulting in 220 sample/normal pairs. Our caller was used to help determine specific forms of alpha thalassemia and estimate copy ratios. We subsequently evaluated the sensitivity and specificity of the annotations, the accuracy of subtype identification, and the precision of copy ratio estimations. As a baseline comparison, we used a previous version of the thalassemia caller used by PiVAT®.

Results. We obtained an overall 88% accuracy rate in annotations with a 96% sensitivity and 93% specificity rate. Furthermore, we found that among the calls that were incorrectly annotated, over 50% of the cases were flagged by the PiVAT® annotation quality score as being poor quality. When low-quality samples are removed, we obtain a specificity rate of 97% while maintaining a sensitivity rate of 95%. By comparison, the previous PIVAT® caller on the same dataset obtained a 67% accuracy rate.

Conclusions. The work on using targeted sequencing to help detecting thalassemia is limited, highlighting the need for improved diagnostic tools. Our thalassemia caller within PiVAT® provides high-accuracy annotations while assessing the uncertainty in these annotations. PiVAT®'s secondary analysis tool demonstrates excellent performance, requiring as little as 10ng of DNA for accurate results. Additionally, the open-source nature of our code ensures accessibility and potential for further development by the research community. This robust and reliable tool enhances the precision and reliability of thalassemia clinical assessment, making a significant contribution to the field.

EXPERIMENT DESIGN

- **Panel coverage:** 131 amplicons covering alpha, beta, delta, gamma and epsilon regions, pseudogenes and control regions
- Median amplicon size of 158bp
- **Clinical samples tested:**
  - 44 positive samples
  - 13 negative
- Sequencing was performed on Illumina's MiSeq™ platform
- Average per amplicon coverage was ~2,800 read pairs
- **Data Analysis:** All the samples were analyzed on Pillar's secondary analysis pipeline, PiVAT®

Sample Type	Sample Zygosity	N
α3.7 del	*Heterozygous*	*31*
α3.7 del	*Homozygous*	*5*
SEA del	*Heterozygous*	*2*
α4.2 del	*Heterozygous*	*2*
FIL del	*Heterozygous*	*1*
α4.2 dup	*Heterozygous*	*1*
Large HBA1/2 dup	*Heterozygous*	*1*
α4.2dup/α3.7del	*Heterozygous*	*1*

Table 1. Breakdown of the expected Thalassemia types of the positive samples tested in this study. Thalassemia types of the samples are shown along with the expected zygosity. Thalassemia types for each sample were previously established using PCR test.

ALGORITHM OVERVIEW

Schematic of PiVAT® 's Algorithm

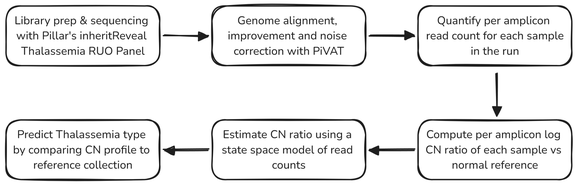


Figure 1. Schematic representation of PiVAT's Thalassemia calling algorithm. The data ingested by PiVAT®'s Thalassemia calling algorithm originate from a targeted sequencing panel. The data undergo several steps for quality control, alignment, and refinement to ensure high quality data are passed to downstream processes. The inputs to the thalassemia calling algorithm are per amplicon and per sample coverages. As with most targeted sequencing data, natural variation in the sequencing coverage exists that does not necessarily reflects the true copy number of the genomic fragment sequenced. Consequently, a CN-free negative control sample is required to normalize the counts. The normalized counts are processed through a state space model to smooth CN estimates and obtain sample quality metric. The smooth CNs are subsequently used to predict the thalassemia type.

Mathematical Modeling of PiVAT®'s State Space Model

State Space Model (Smoothing)

- State space model expects copy ratios between adjacent amplicons to differ with probability p
- These abrupt changes in copy ratio have variance σ^2

$$p(x_1) = \mathcal{N}(0,1)$$
$$p(x_{j+1}|x_j) = p\mathcal{N}(x_j, \epsilon) + (1-p)\mathcal{N}(x_j, \sigma^2)$$
$$p(y_j|x_j) = Laplace(x_j, b)$$

Thalassemia Profiling

- We are given a list of profiles $\{d_i\}$ defined by $E[x|D = d_i]$ (the expected copy ratio per amplicon)
- Optimal decision: profile with maximum posterior density

$$\hat{d} = \underset{d}{\operatorname{argmax}} \int p(D = d|x)p(x|y)dx$$

- Simply reuse samples from particle filter to approximate

Bayesian Evidence-based Sample Quality Estimate

- Use the median to ignore jumps, compare between samples

$$QC = Median\left(p(y_j|y_{1:j-1})\right)$$

RESULTS

Overall Results Summary

		Predicted			
		No QC		**QC**	
		Negative	**Positive**	**Negative**	**Positive**
Actual	**Negative**	31	13	29	5
	Positive	1	175	1	168
	Accuracy	*88.7%*		*93.7%*	

Figure 2. Confusion matrix with and without using sample QC based filtering. We evaluate the detection of a variant both with and without the QC step defined by the third equation. Without this QC step, we observe 13 false positives. When filtering is applied, this drops to 5. We also evaluate our accuracy in predicting the exact thalassemia profile, which without QC step is 88.7% accuracy, which rises to 93.7% accuracy with QC.

RESULTS

Impact of Smoothing on Estimated Copy Number Ratio

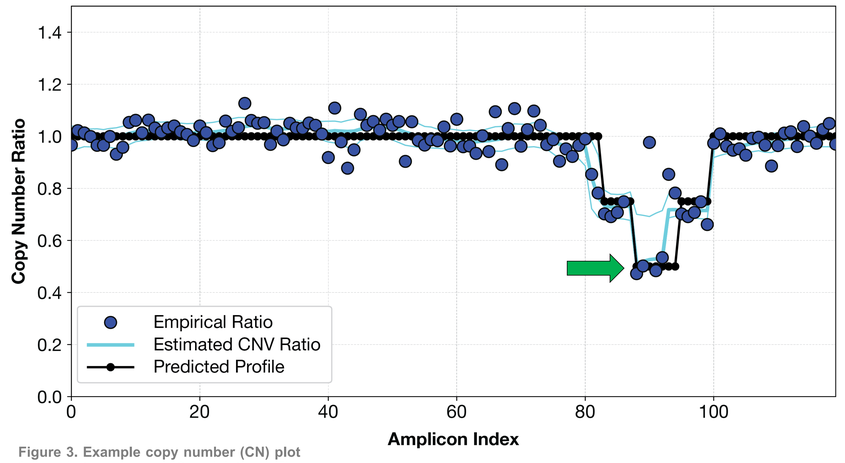
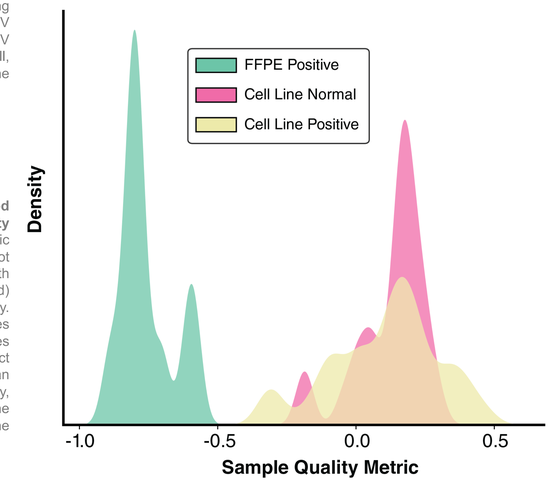


Figure 3. Example copy number (CN) plot for a heterozygous α3.7 deletion. The figure shows the empirical CN ratios across all amplicons for a α3.7 deletion sample. The empirical ratios are smoothed out using PiVAT®'s space model to estimate the CNV ratio shown in cyan line. This estimated CNV ratio is used to identify the thalassemia call, matching the expected deletion shown by the green arrow.

Bayesian Evidence-based Sample Quality Estimation

Figure 4. Stratification of sample based on a Bayesian evidence-based quality metric. Demonstration of the quality metric in separating samples whose profile do not match those of the normals. Samples with quality metric three standard deviation (3sd) below the mean are rejected as poor quality. In this example, all FFPE positive samples were compared to cell line normal samples and were deemed poor in quality. The impact of FFPE damage makes a cell line normal an inadequate comparator and consequently, makes CNV prediction inaccurate. Cell line positives samples share similar profile to the normal and are all above our 3sd cutoff.



CONCLUSIONS

- Pillar's inheritReveal™ Thalassemia RUO panel can accurately help profile thalassemia in clinical samples using a targeted sequencing-based approach.
- Here we present a new machine learning approach for helping profile thalassemia types in targeted sequencing dataset.
- We also implement a Bayesian evidence-based quality metric to improve overall profiling and prediction accuracy.
- Larger studies would be needed to establish the clinical utility of this approach.