# General validation framework using semi-supervised learning on complex cfDNA clinical assays

Counsyl

Kevin R. Haas, PhD; Kyle A. Beauchamp, PhD; Jeff Tratner, BA; Kevin D'Auria, PhD;
Chuba Oyolu, PhD; Carrie Haverty, MS LCGC; Dale Muzzey, PhD

*South San Francisco, California*

## Introduction

With advancements that enable detection of a greater range and complexity of genetic variation, there are increasing challenges in effectively validating clinical genomic tests and proving their analytical performance. This difficulty is especially acute in applications of the rapidly growing field of cell-free DNA sequencing, such as noninvasive prenatal screening (NIPS) and circulating-tumor DNA analyses, where the aim is to detect miniscule signal from placental or tumor DNA, respectively.

## Challenges of conventional concordance evaluation of bioinformatics pipeline

In a typical bioinformatics pipeline, a variant is called positive if the confidence score (e.g., z-score, likelihood) exceeds established thresholds, which can vary depending on the type of variant, assay chemistry, and calling algorithm. Tuning this threshold and then evaluating sensitivity and specificity can be difficult (1) when reference samples with consensus genotypes are scarce and (2) when the underlying biology of the assay necessarily causes the signal of positive samples to approach the limits of detection.

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

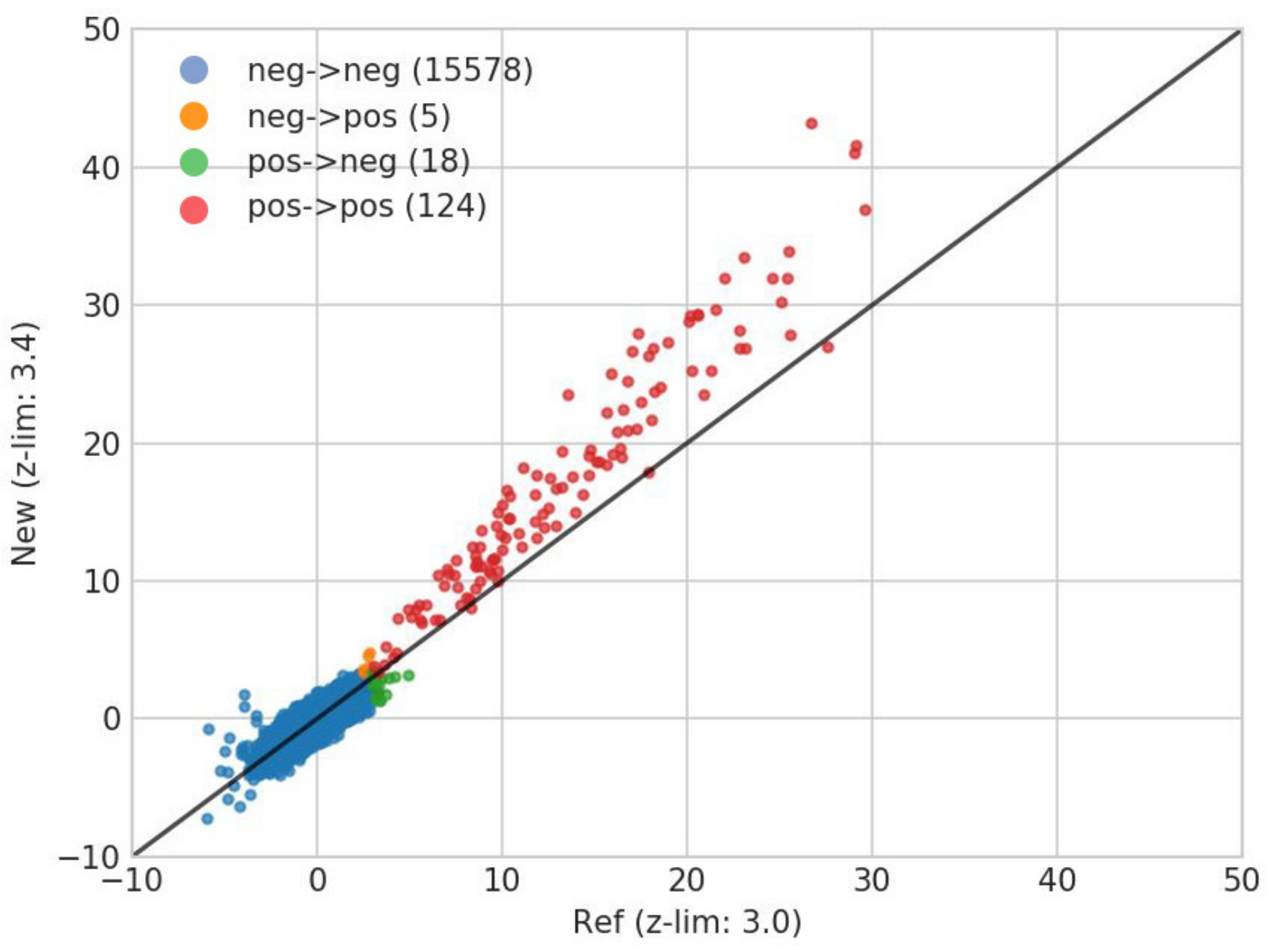


**Figure 1: Comparison of chromosome 21 z-score from NIPS calling algorithms.**
For NIPS, the z-score for a region assesses the normalized enrichment in mapped WGS reads. 15,725 Counsyl production NIPS samples were analyzed with both a reference and a new bioinformatics algorithm pipeline. The vast majority of these samples have no reported clinical outcomes, so true caller performance cannot be directly obtained. Therefore, candidate z-score limits of 3.0 and 3.4 were applied to the reference and new methods, respectively, to attain hypothetical concordance of called trisomy 21 between the two methods.

| Chromosome 21 | True Positive | True Negative |
|---|---|---|
| Called Positive | 125<br>(124-126) | 14<br>(8-22) |
| Called Negative | 1<br>(0.5-2) | 15,585<br>(15-577-15,591) |

**Table 1: Discordances associated with pipeline changes occur with similar frequency as discordant clinical outcomes in meta-analysis.**
Using the reported clinical sensitivity 99.2% (98.5%-99.6%) and specificity 99.91% (99.86%-99.95%) from Gil et. al.[1], one would expect from a high risk population[2]—i.e., with 1/125 incidence of trisomy 21—the following clinical outcomes for a similar hypothetical cohort of 15,725 samples. This illustrates the magnitude of discordance from bioinformatics pipeline changes is commensurate and inherent to the statistical performance of the assay.

**REFERENCES 1.** Gil MM, Akolekar R, Quezada MS, Bregant B, Nicolaides KH. Analysis of cell-free DNA in maternal blood in screening for aneuploidies: meta-analysis. Fetal diagnosis and therapy 2014;35(3):156-73. **2.** Snijders RJM, Sundberg K, Holzgreve W, Henry G, Nicolaides KH. Maternal age and gestation-specific risk for trisomy 21. Ultrasound Obstet Gynecol 1999;13:167-70. **3.** Kevin R. Haas PhD, Kevin D'Auria PhD, Jeff Tratner BA, Chuba Oyolu PhD, Carrie Haverty MS LCGC, Dale Muzzey PhD "Accurate Fetal Fraction from NIPS using Whole Genome Sequencing" ISPD 2017. **4.** Patil A, Huard D, Fonnesbeck CJ. PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software. 2010;35(4):1-81.

## MCMC posterior prediction of Bayesian graphical model of assay performance

To overcome the challenge of assessing sensitivity and specificity, we have developed a Bayesian graphical modeling approach capable of deconvoluting and parameterizing the negative and positive distributions from empirical data of unlabeled samples. One begins by postulating the interrelation between the latent distributions for variant incidence, allele fraction, and sample classification with the observed confidence score.

$$p(FF|FF_{true}) \sim \mathcal{N}(\mu = FF_{true}, \sigma = \sigma_{FF})$$

$$p(FF) \sim Beta(\alpha_{ff}, \beta_{ff})$$

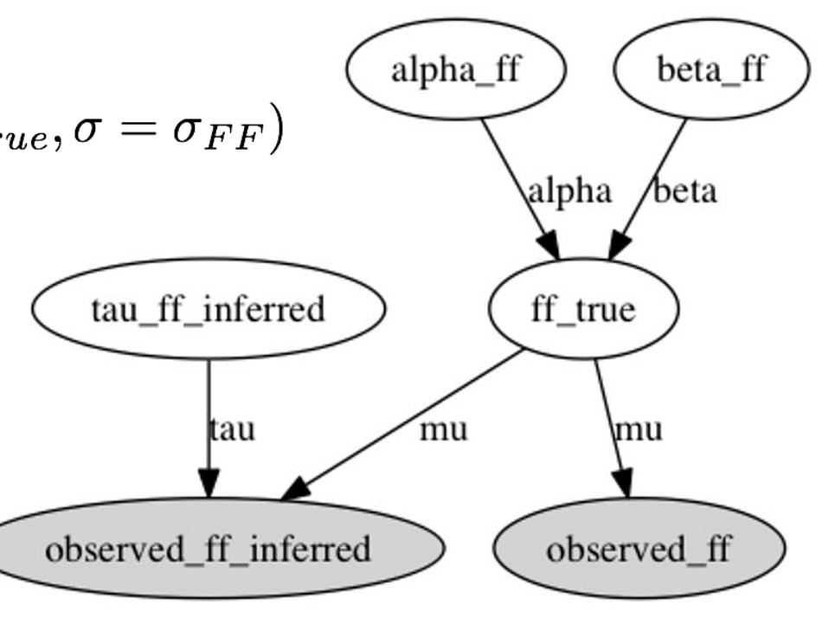
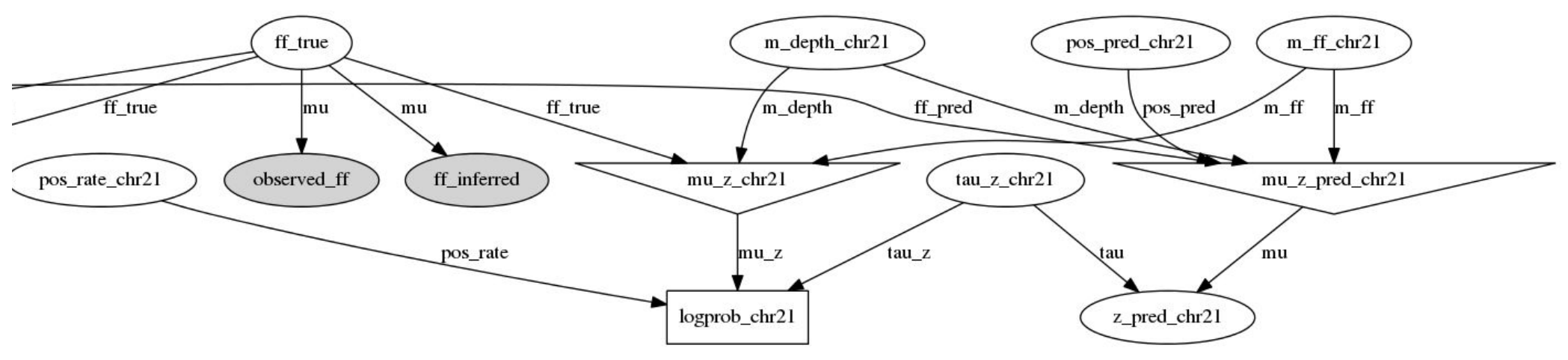


**Figure 2: Bayesian graphical model for fetal fraction.**
Diagram of the statistical relationship between the observed fetal fraction (*FF*) from chrY enrichment for male fetuses, which is statistically dependent on the unobserved true fetal fraction (*FF$_{true}$*) through a normal distribution. The true fetal fraction is itself drawn from a beta prior. Additionally, one may calculate and relate an inferred (*FF$_{inferred}$*)[3] which is observed for both male and female fetuses.



$$p(z_i|\chi_i) \sim \begin{cases} \mathcal{N}(\mu = \mu_i^+, \sigma = \sigma_i) & \chi_i = trisomy \\ \mathcal{N}(\mu = 0, \sigma = \sigma_i) & \chi_i = wildtype \\ 1.0 & \chi_i = no-call \end{cases}$$

$$\mu_i^+ = m_{FF} \log_2(1.0 + FF_{true}/2.0) \frac{\sqrt{d}}{\sqrt{d} + m_{depth}}$$

**Figure 3: Bayesian graphical model for z-scores from NIPS calling.**
The mean of z-scores for positive samples is assumed to increase with fetal fraction or minor allele fraction according to the slope mFF and increase with normalized sequencing depth with the adjustable inflection point of mdepth. The observed z-score is included in the calculation of the log probability from a mixture model of the mostly unknown trisomy, wildtype, and no-call category for the sample. The class prevalence is drawn from a Dirichlet distribution. Additionally the model depicts predicted z-scores with explicitly labeled positive or negative category predictions.

Inference of the posterior predictive distribution—which assigns probabilistic outcome labels to the previously unlabeled data—is performed using Markov Chain Monte Carlo (MCMC)[4]. We then parametrically scan a range of classification thresholds and calculate sensitivity and specificity to construct a ROC curve with confidence intervals given by bootstrap sampling.

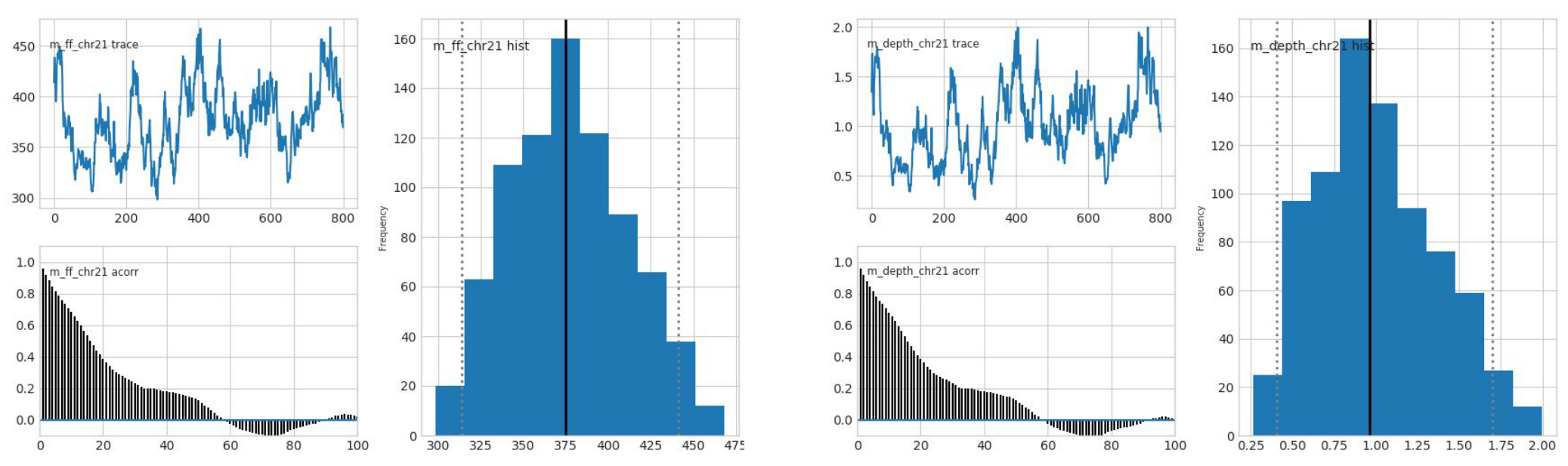


**Figure 4: MCMC sampling of Bayesian graphical model.**
Inference of the model on cohort data is performed using MCMC sampling to explore the full posterior distributions for the model parameters of trisomy pos_rate, *m$_{FF}$* (left panel) and *m$_{depth}$* (right panel). The MCMC sampling allows capture of the trajectories of the parameters over the course of sampling, as well as histograms of the data.

## Results for illustrative example from NIPS algorithm advancement

We present an illustrative example of how this statistical approach has been leveraged to evaluate algorithmic enhancements to our NIPS offering using 15,725 random anonymized clinical samples, including >500 positives. From the ROC curve, we determined that sensitivity increased significantly, with a ~3x reduction in the false-negative rate, while maintaining specificity.

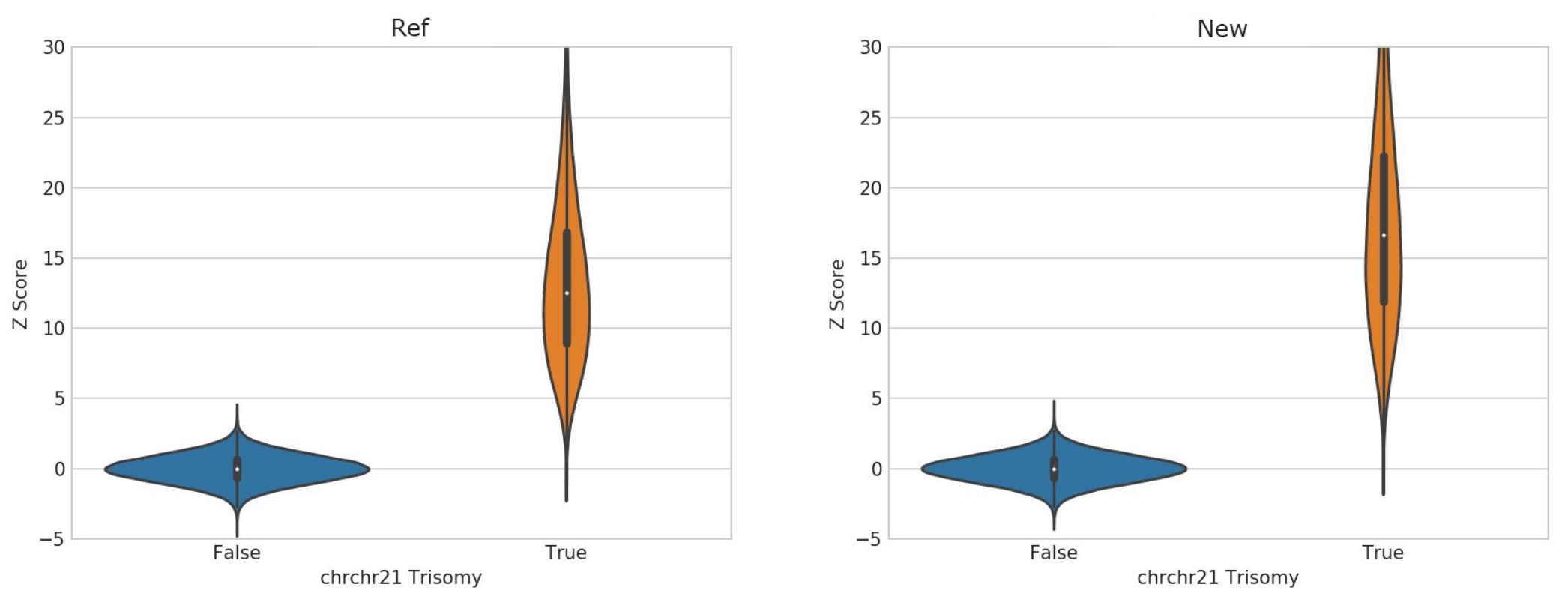


**Figure 5: Posterior predictive z-score distributions for chromosome 21.**
MCMC predicted data by sampling fetal fraction from the trained prior, positive/negative labels from the trained positive rate, and finally z-scores given the learned dependence on fetal fraction and depth. Using these posterior predictive samples, we reconstruct the estimated distributions for false/true trisomy 21 shown for both the reference caller and the new caller. The negative distributions are essentially unchanged while the positive distribution is shifted to higher z-scores indicating improved classifier performance.

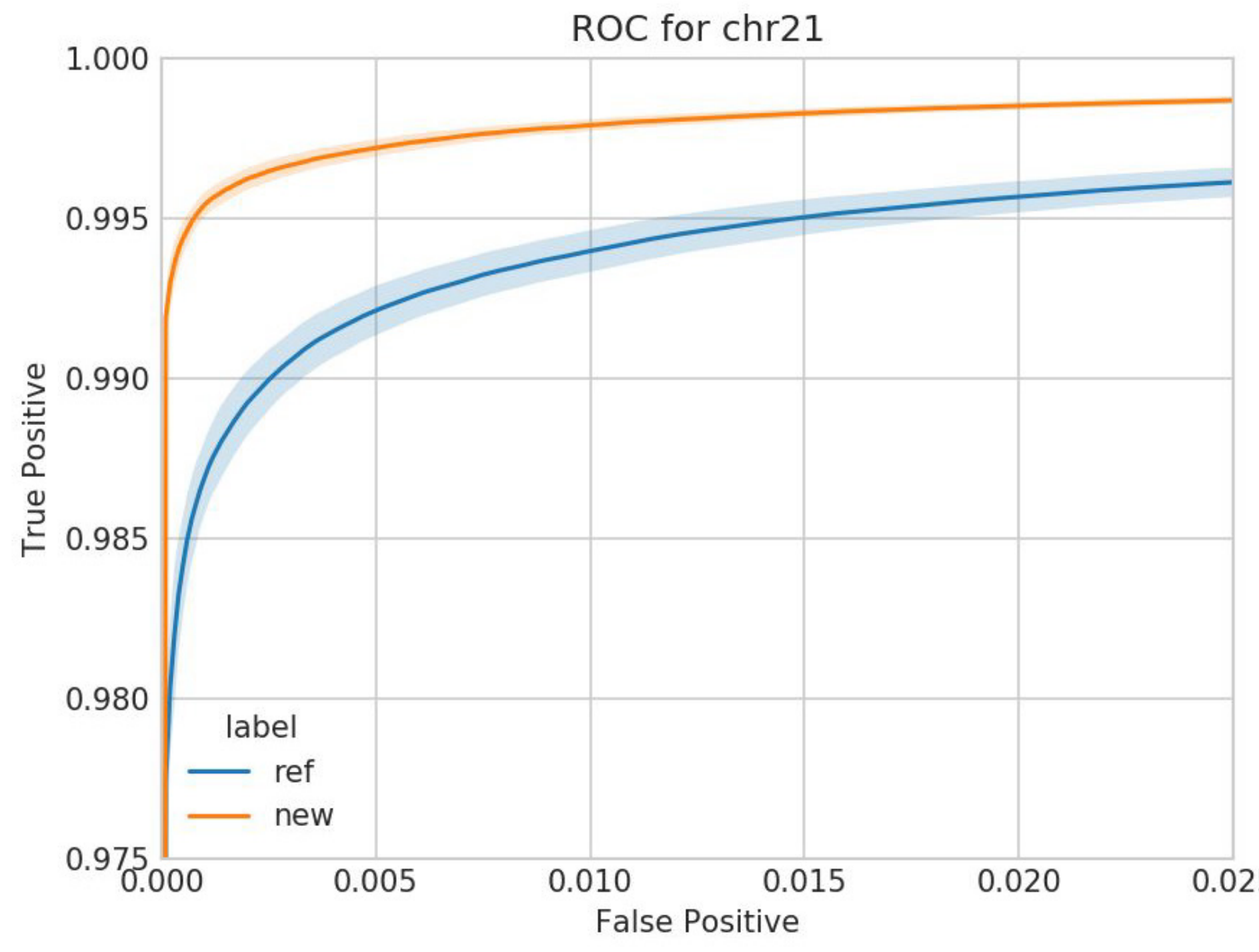


**Figure 6: Estimated ROC curves from MCMC sampling of chromosome 21.**
The quantitative relationship between true- and false-positive rates can be calculated by parametrically scanning the z-score threshold—which delineates positive and negative samples—over the posterior predictive distributions of z-scores. The portion of each distribution that lies above or below this threshold indicates the specificity and sensitivity, respectively. The error bounds (shaded) represent the standard deviation in the ROC estimate from 10-fold bootstrap sampling of the original dataset.

## Conclusion

Semi-supervised Bayesian inference allows assessment of analytical performance from a real-world collection of test samples for cutting-edge genetic assays. Furthermore, we explored the relationship between sequencing depth, minor allele fraction and confidence score to decipher the limits of detection.

**View all posters and research at**
research.counsyl.com

Q506E V1 10/17