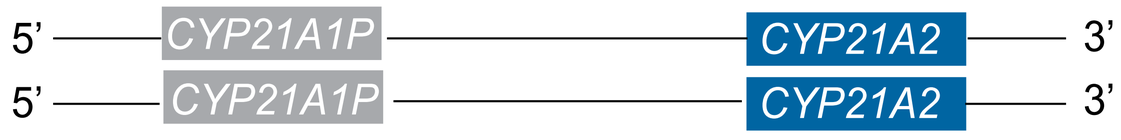
A deep learning model for accurate variant calling congenital adrenal hyperplasia

Sun Hae Hong, PhD; Adithya Ganesh; Kyle A. Beauchamp, PhD; Dale Muzzey, PhD; Kevin R. Haas, PhD
Myriad Women's Health

BACKGROUND

- Congenital adrenal hyperplasia (CAH) is an autosomal recessive disease that impairs steroidogenesis.
- Mutations in *CYP21A2* account for a large fraction of CAH cases. *CYP21A2* and a pseudogene *CYP21A1P* have high sequence identity (Fig 1).
- Variant calling in *CYP21A2* is technically challenging due to frequent and complex gene rearrangements with *CYP21A1P*.
- We developed an enhanced deep learning model (deepCAH) which introduces additional features and class labels to improve CAH variant calling.

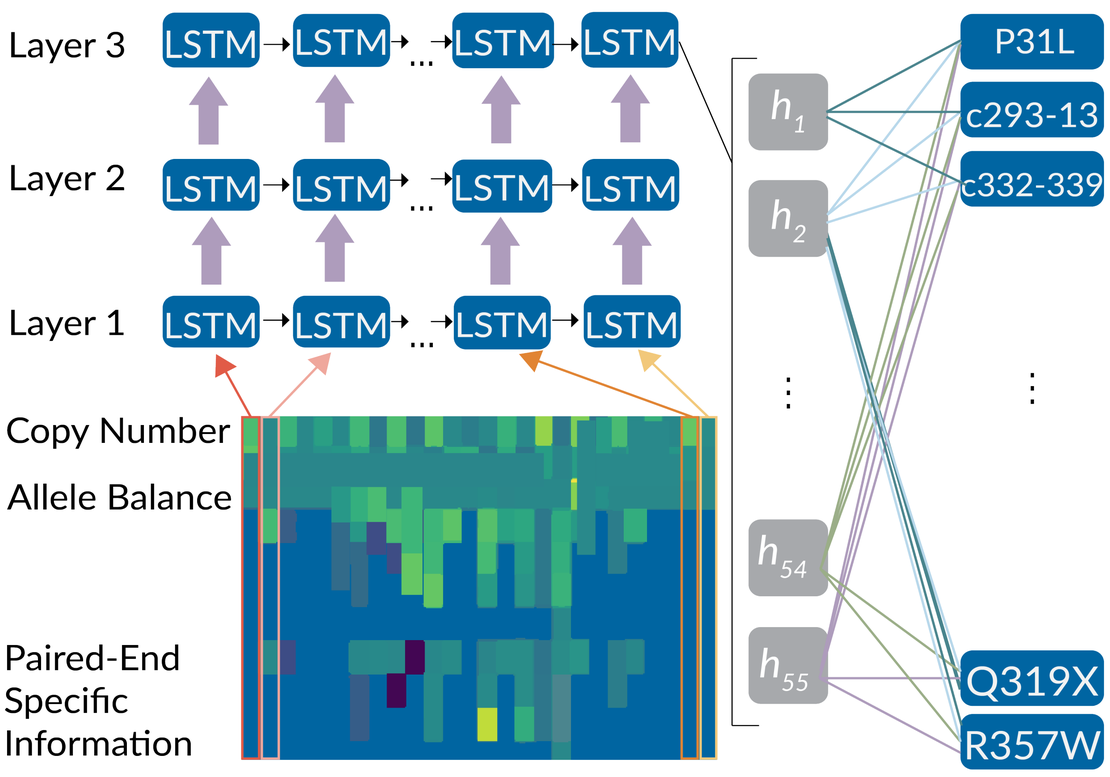
Figure 1. Gene structure of CAH in a healthy individual.



METHODS

- Our original deepCAH model utilized 2 sets of features:
  - Allele-balance measurements of benign and deleterious alleles of interest.
  - Copy-number measurements consisting of normalized sequencing depths at differentiable loci on *CYP21A2* and *CYP21A1P*.
- The extended deepCAH model includes novel features—paired end-specific read contributions, which are often examined during manual call review to assess the fidelity of read contribution.
- In addition to 11 variants, the extended deepCAH model included a new variant indicating a deleterious mutation (Q319X) is in cis with gene duplication.
- A cohort of >37,000 research-allowed samples — between March 2019 and May 2019 was split into training (80%) and test sets (20%).
- We employed a 3-layer recurrent neural network comprised of long short-term memory (LSTM) cells and a weighted cross-entropy loss function (Fig 2).
- The model was implemented in TensorFlow and trained using the Adam optimizer.

Figure 2. deepCAH architecture. The model has 3 layers of LSTMs. Features corresponding to genomic location were input to model sequentially. The output layer is fully connected.



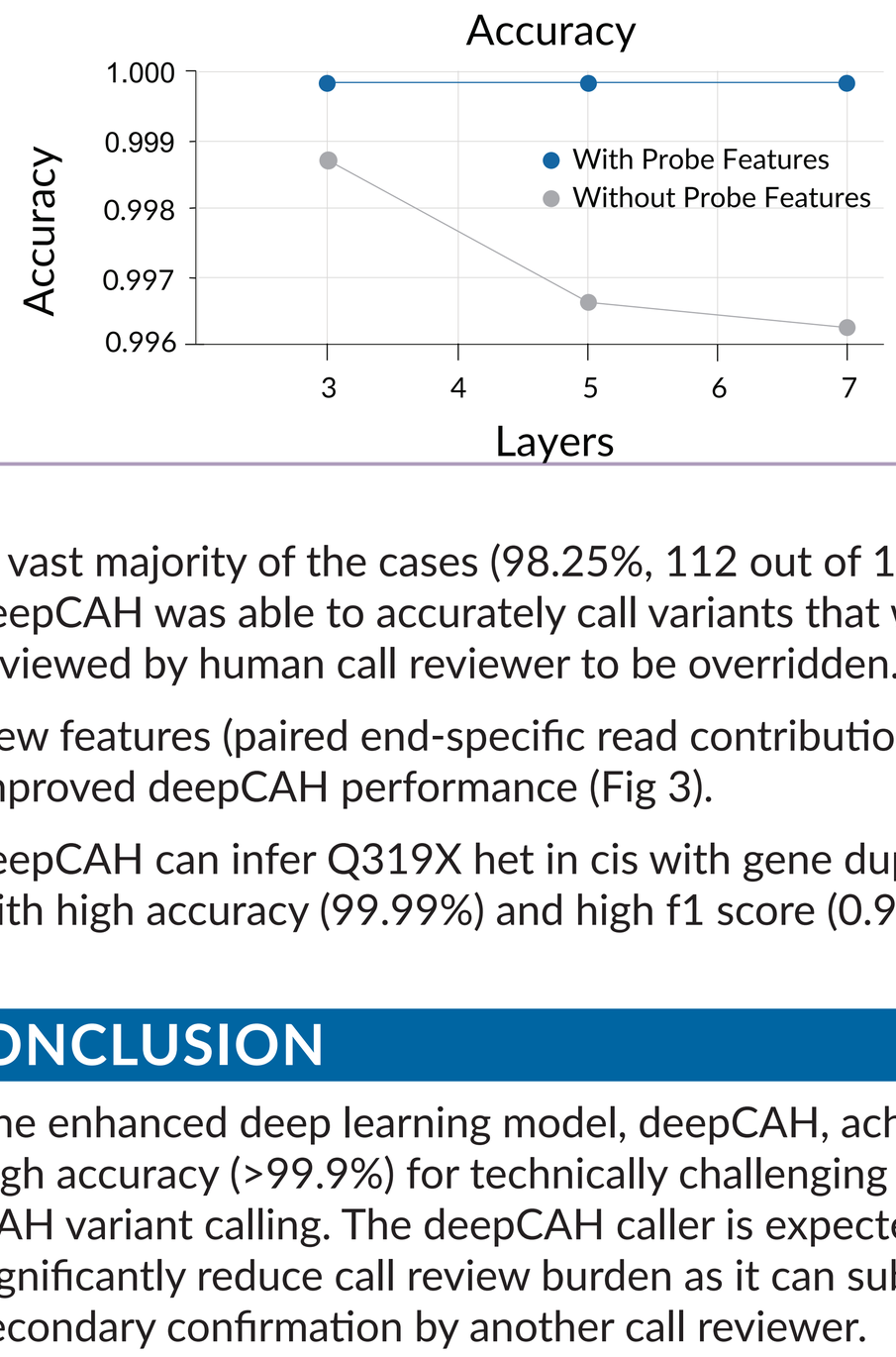
RESULTS

Table 1. Confusion matrix of SNP variants in test data.

	True Call	
deepCAH	**Negative**	**Positive**
Negative	82481	2
Positive	4	761

- Considering the human-reviewed calls as ground truth, the extended deepCAH showed accuracy of 99.99% and f1 score of 0.9958 in test set. Out of 83248 SNP calls, there was only 2 false positives and 4 false negatives (Table 1).

Figure 3. New features (paired end-specific read contribution) improved deepCAH performance.



- In vast majority of the cases (98.25%, 112 out of 114 calls), deepCAH was able to accurately call variants that were reviewed by human call reviewer to be overridden.
- New features (paired end-specific read contribution) improved deepCAH performance (Fig 3).
- deepCAH can infer Q319X het in cis with gene duplication with high accuracy (99.99%) and high f1 score (0.9961).

CONCLUSION

- The enhanced deep learning model, deepCAH, achieved high accuracy (>99.9%) for technically challenging CAH variant calling. The deepCAH caller is expected to significantly reduce call review burden as it can substitute secondary confirmation by another call reviewer.

All posters available at research.myriadwomenshealth.com
Presented at ASHG on October 16, 2019