A Deep Learning Framework for Interpreting Repetitive DNA Sequence in Heritable Disease

Lando Ringel, Julie Thibert, Abhinay Reddy, Elliot Hallmark, Jacob Ashton, Brian C Haynes, Sarah Statt and Jessica L Larson
Asuragen, Inc., Austin, TX

Summary

- Pathogenic STRs elude interpretation by NGS due to their length and low sequence complexity.
- We present DeepNet*, a prototype deep learning approach to enable the accurate genotyping of short tandem repeat (STR) variants.
- The model achieved >99% sensitivity and PPV the genotype level and >99% accuracy at the sample category level in independent evaluation datasets across multiple capillary electrophoresis (CE) platforms.

Introduction

Advances in PCR/CE technologies have enabled the analysis of STR DNA fragment size through electropherograms (traces) with visually identifiable peaks that are translatable into corresponding genotypes. Existing approaches for PCR/CE peak discernment require manual inspection or heuristic algorithms tailored to assay- and/or instrument-specific signal idiosyncrasies. We present a push-button analysis tool (DeepNet) with a convolutional neural network (CNN) that circumvents these limitations, while robustly detecting and interpreting STR alleles relevant to clinical research and patient testing.

Materials and Methods

Biological specimens from the Coriell NINDS repository with known hexanucleotide repeat lengths in *C9orf72*, an ALS-associated gene (NINDS; Figure 1), were collected and analyzed with the AmplideX® PCR/CE *C9orf72* Kit† on two Genetic Analyzer platforms. Trained operators adjudicated and confirmed expected allele lengths through manual analysis of electropherograms. The CE traces were portioned into training and testing cohorts for model development and evaluation. The trained CNN (Figure 2) was evaluated on the Coriell test set and an independent external residual clinical specimen cohort.

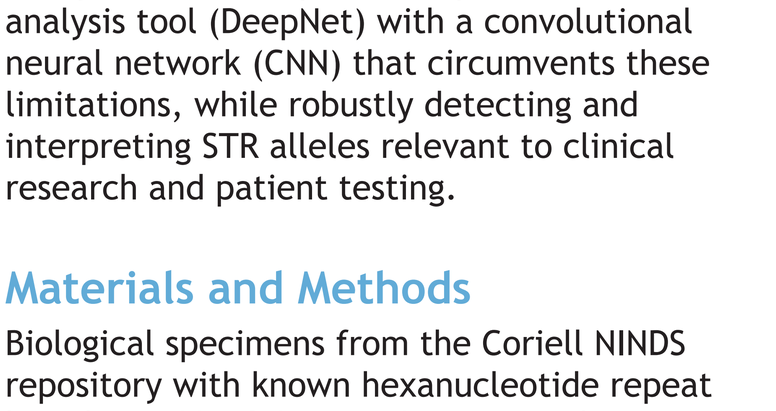

Figure 1. Representation of **A)** the 11 exons in the *C9orf72* gene with the location of the G_4C_2 repeat denoted with vertical lines and **B)** 3-primer gene-specific (GS) and repeat primed (RP) design of the AmplideX PCR/CE *C9orf72* Kit. Adapted from Bram et al. 2018, *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*.

*Proof-of-concept data only. Future availability and performance cannot be ensured.
†Research Use Only. Not for use in diagnostic procedures.

Presented at ASHG 2019

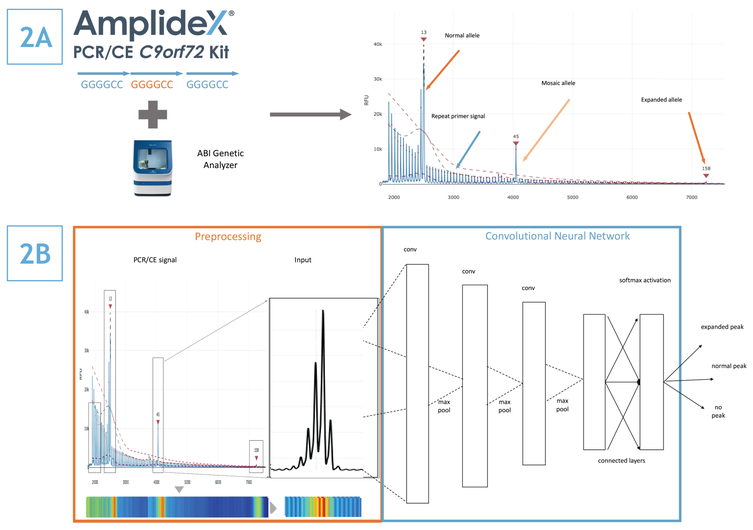

Figure 2. A) All specimens were analyzed with the AmplideX Kit, resulting in unique traces. Representative 3-primer GS/RP-PCR/CE profile for an expanded specimen (ND06769) with GS peaks (orange arrows) and RP-PCR profile (blue arrow). **B)** We reframed PCR/CE analysis as a computer-vision image classification problem by representing the multi-channel signal as a 1D image. Potential allelic peak regions were automatically selected and classified as allelic peaks or background signal with multiple convolutional layers and pooling operations using Keras and TensorFlow. Repeat-primed patterns at the end of the trace were associated with expansion events.

Results

Table 1. Categorical performance metrics for DeepNet in the training **(top)** and test **(bottom)** cohorts. The CNN achieved perfect accuracy at the sample category level (100%) in both data sets.

AmplideX DeepNet (training cohort)	Expected Genotype			
Genotype Category # GGGGCC	Normal ≤20	Intermediate 21-29	Expanded ≥30	Total
Normal ≤20	431	0	0	431
Intermediate 21-29	0	11	0	11
Expanded ≥30	0	0	110	110
Total	431	11	110	**552**

AmplideX DeepNet (testing cohort)	Expected Genotype			
Genotype Category # GGGGCC	Normal ≤20	Intermediate 21-29	Expanded ≥30	Total
Normal ≤20	109	0	0	109
Intermediate 21-29	0	2	0	2
Expanded ≥30	0	0	39	39
Total	109	2	39	**150**

Table 2. Allele-level performance metrics for the deep learning genotyping algorithm in Coriell (enriched for mosaicism) and residual clinical sample training cohorts. The model only missed one allele out of 1122. Mosaic alleles were excluded from false-positive (FP) calls.

	Coriell training cohort (3500xL)	Coriell training cohort (3130xL)	Clinical training cohort (3130xL)	Training totals
# of specimens	469	34	46	552
Expected # of peaks	938	92	92	1122
TP	938	92	91	1121
FN	0	0	1	1
FP	0	0	1	1
Sensitivity	100%	100%	98.9%	**99.9%**
PPV	100%	100%	98.9%	**99.9%**

Table 3. Allele-level performance metrics for the deep learning genotyping algorithm in Coriell and residual clinical sample independent testing cohorts. The model only missed two alleles out of 300; a 6|12 was called an 8|14, resulting in two FPs and false-negative (FN) calls at the peak level. Mosaic alleles were excluded from FP calls.

	Coriell testing cohort (3500xL)	Coriell testing cohort (3130xL)	Training totals
# of specimens	100	50	150
Expected # of peaks	200	100	300
TP	198	100	298
FN	2	0	2
FP	2	0	2
Sensitivity	99.0%	100%	**99.3%**
PPV	99.0%	100%	**99.3%**

Table 4. The limit of detection for expanded mosaic peaks in the background of normal *C9orf72* alleles is 20% (empirically determined).

	Number of expected peaks	Number correctly called	Percentage correctly called
40% allele frequency	12	12	100%
20% allele frequency	12	12	100%
10% allele frequency	12	9	75%

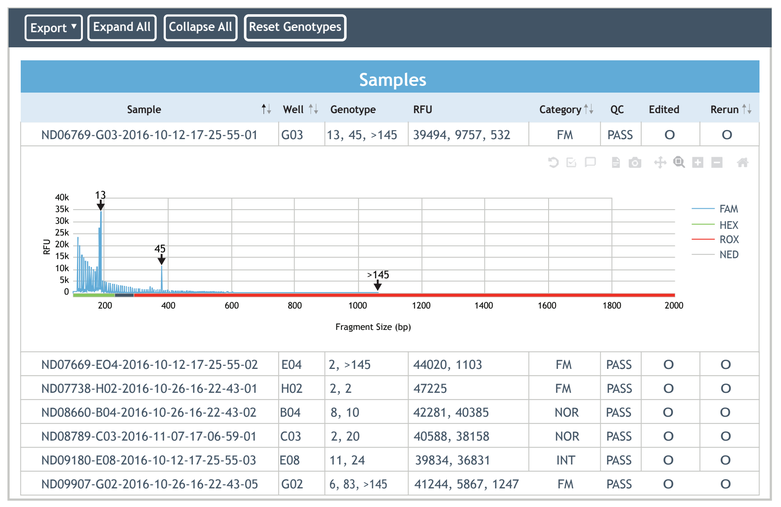

Figure 3. The DeepNet pipeline for *C9orf72* was deployed as a standalone analysis module in the AmplideX Reporter software platform. The module enables rapid (<3 min per 100 samples) reporting, automated QC checks, and a web interface for data review.

Conclusions

- This deep-learning method was able to accurately distinguish genotype-relevant signal from background noise in a *C9orf72* STR PCR/CE assay.
- The modularity and extensibility of our pipeline may spur the rapid development of new designs to accelerate clinical research and precision diagnostics in STR disorders and other PCR/CE applications.

Asuragen®
www.asuragen.com