A Deep Learning-Powered Genotyping System for *C9orf72* Hexanucleotide Repeat Expansions Enables High Throughput Genetic Analysis

Ryan Routsong, Adrian Gonzalez, Lando Ringel, Jacob Ashton, Sarah N Statt, Gary J Latham and Brian C Haynes

Asuragen, a Bio-Techne brand, Austin, TX

Summary

- Accurate quantification of hexanucleotide repeats in the *C9orf72* gene is essential to understanding genotype-phenotype relationships in gene-associated disorders, such as ALS and FTD.
- An automated analysis solution for the AmplideX® PCR/CE *C9orf72* Kit†, the AmplideX PCR/CE *C9orf72* Analysis Module*, achieves accuracy on par with expert manual analysis.
- The peak analysis software utilizes a deep convolutional neural network (CNN) trained on large datasets to predict genotypes with high accuracy, resulting in an exponential reduction in analysis time while avoiding common peak annotation errors.

Introduction

Hexanucleotide expansions in the open reading frame 72 of human chromosome 9 (*C9orf72*) are a principal genetic driver of ALS-Frontotemporal spectrum disorder. Genotyping of *C9orf72* by PCR/Capillary Electrophoresis (CE) currently requires a significant amount of manual analysis time and may still yield inconsistencies between trained technicians. We developed a fully automated deep learning approach that achieves human-level performance while significantly reducing the time required for manual interpretation.

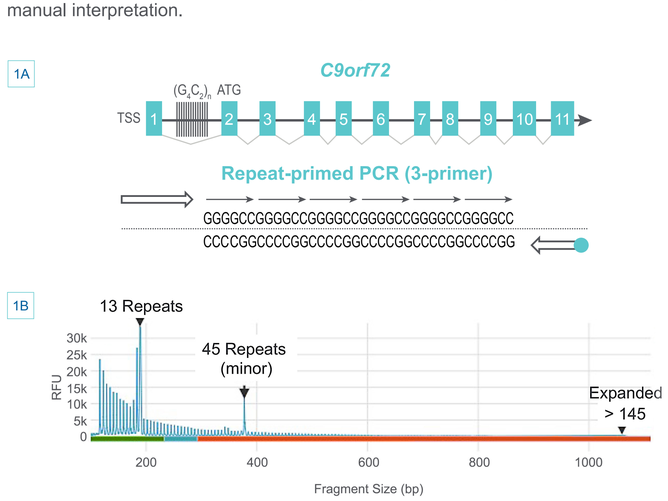

Figure 1. Schematic Representation of the *C9orf72* Gene Structure Showing the Predicted 11 Exons (Boxes) and Location of the Intronic Hexanucleotide Repeat Expansion (Vertical Lines). AmplideX PCR/CE *C9orf72* Kit† 3-primer FAM-labeled repeat-primed PCR software output. Coriell sample ND06769 displayed with gene specific repeat peaks labeled (13, >145, and a 45-repeat minor allele) and hexanucleotide repeats producing the "sawtooth" repeat pattern; genotype categories identified by the following colors: green = normal (0-19 repeats), blue = intermediate (20-29 repeats), and orange = expanded (≥30 repeats).

†Research Use Only – Not for Use in Diagnostic Procedures
*This product is under development. Future availability and performance to be determined.
Presented at AMP 2021

Materials and Methods

DNA was isolated from blood or acquired from Coriell Institute for Medical Research cell lines across multiple cohorts including the National Institute of Neurological Disorders and Stroke (NINDS) ALS sample set.[1] Over 1500 electropherograms were generated by AmplideX PCR/CE *C9orf72* Kit† (Asuragen) across three CE instruments: 3130*xl*, 3730*xl*, and 3500xL (Thermo Fisher). A subset of data was used to train a convolutional neural network (CNN). The algorithm evaluates each region of the trace to identify genotype peaks, further determining sizing category and sample QC based on overall interpretability of the sample. The CNN genotyping algorithm and QC logic was packaged into push-button reporting software for use with the PCR/CE assay.

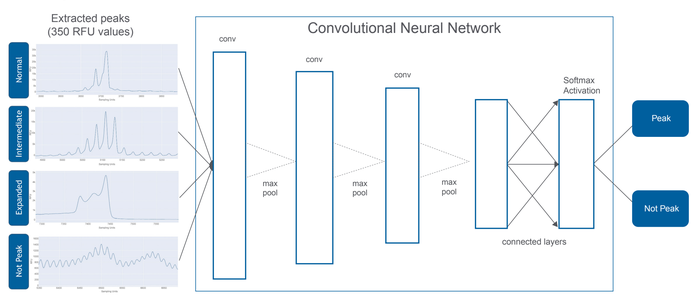

Figure 2. Convolutional Neural Network (CNN) Architecture of the *C9orf72* Peak Correction Model. A classic series of convolutional layers, followed by max pooling layers terminating into three fully connected layers with Rectified Linear Unit (ReLU) activation function and output layer with a softmax activation function. The network outputs a probability of a region containing a genotype peak. A multi-channel, windowed representation of candidate regions in the CE trace is used as input features for the network.

Results

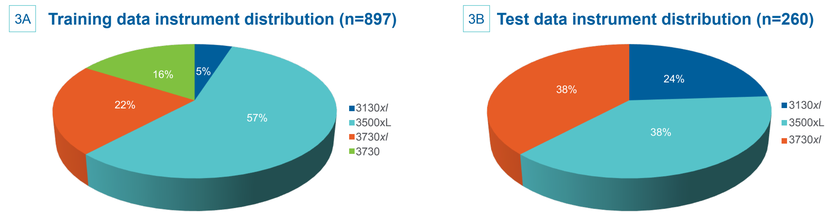

Figure 3. Instrument Platform and Configuration Distribution of Training and Testing Data in the *C9orf72* Analysis Algorithm. Distribution of instruments and platforms are semi-balanced in the training **A)** and test **B)** data sets due to inclusion of important edge cases and normalized representation of otherwise benign data points. The training set **A)** shows a majority of 3500 platform data semi-balanced expression of other platforms, while the testing dataset **B)** shows a well-balanced mix of the different platforms.

1A

EXPECTED \ PREDICTED	NORMAL	INTERMEDIATE	EXPANDED
NORMAL	215	0	0
INTERMEDIATE	0	3	0
EXPANDED	1	0	41

1B

EXPECTED \ PREDICTED	0-19	20-29	≥30
0-19	484	0	0
20-29	0	4	0
≥30	1	0	22

Table 1. Genotyping Accuracy on an Independent Test Cohort of 260 Samples. Sample genotype categorization **A)** and genotype-peak-level sizing accuracy **B)** exceeded 99% accuracy according to established categorical size ranges[1]: normal (<20 repeats), intermediate (20-29 repeats), and expanded (≥30 repeats). The only discordant sample was an expected expanded (10,162) classified as normal (10,10). All peaks called in the range of 0-30 repeats were sized within ±1 repeat of the expected allele size.

Samples

Sample	Well	Genotype	RFU	Category	QC	Edited	Rerun
ND04056-A01-2016-10-26-16-22-43-01	A01	2, 10	43607, 40124	NOR	PASS	○	○
ND07489-B02-2016-11-07-17-06-59-01	B02	2, >145	41605, 293	FM	PASS	○	○
ND07920-H04-2016-08-29-14-55-51-04	H04	2, >145	42447, 2039	FM	PASS	○	○
ND08078-F05-2016-10-12-17-25-55-02	F05	6, >145	40566, 940	FM	PASS	○	○
ND08554-H02-2016-11-09-16-32-51-01	H02	5, >145	40143, 270	FM	PASS	○	○

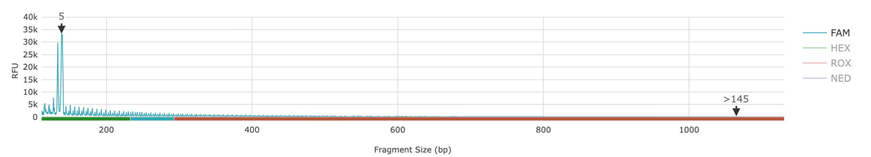

Figure 4. The Genotyping Algorithm of the AmplideX PCR/CE *C9orf72* Analysis Module* Follows a Robust Algorithm for Quality Checking Which Involves Signal-related Quality Checks Against Saturation, High Signal Magnitude, and Quality of the ROX Ladder Signal. The quality control algorithm checks various signal conditions to determine if output genotypes are interpretable. Other checks occurs to ensure the genotype is feasible (genotype QC; i.e., no gene-specific peak was found) and the control is predicted as expected (control QC). Figure 4 shows a screenshot of the AmplideX® Reporter *C9orf72* display, highlighting the results view of the software.

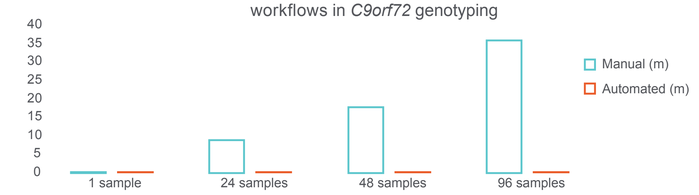

Figure 5. Internal Timed Workflow Studies Show That There is an Exponential Time Savings When Using the AmplideX PCR/CE *C9orf72* Analysis Module* Genotyping Algorithm Compared with Manual Microsoft Excel® Macro Workflows. Manually assessed samples were collected and processed as part of the design verification testing phase of assay development for the AmplideX PCR/CE *C9orf72* Kit† for a twenty-four-sample cohort and then extrapolated for larger or small cohorts. The *C9orf72* genotyping algorithm was timed for each of the 4 different sample sizes.

Conclusions

- The AmplideX PCR/CE *C9orf72* Analysis Module* provides an automated workflow that is as accurate as manual operators but takes a fraction of the processing time.
- Specifically, the AmplideX PCR/CE *C9orf72* Analysis Module* accurately categorized 99.6% of samples from a 260-member test cohort while generating repeat genotypes from a 96-well plate of samples in less than 10 seconds - at least 200 times faster than manual operators.
- In ongoing work, six international and domestic laboratories are evaluating the AmplideX PCR/CE *C9orf72* Analysis Module*. The results from this multi-laboratory assessment can help harden its analytical models and provide independent validation of the assay.

References

1. Bram, Eran, et al. "Comprehensive genotyping of the *C9orf72* hexanucleotide repeat region in 2095 ALS samples from the NINDS collection using a two-mode, long-read PCR assay." Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 20.1-2 (2019): 107-114.

Asuragen
a bio-techne brand