Application of ONT long read sequencing to confirm microdeletions and microduplications in a clinical setting.

Akash Kumar[1], Stephanie Greer[1], Jacquelin Botello[1], Donna Hongo[1], Brynn Levy[1], Premal Shah[1], Matt Rabinowitz[1,2], Kate Im[1] [1]MyOme, Inc., Menlo Park, CA; [2]Natera, Inc., San Carlos, CA

BACKGROUND

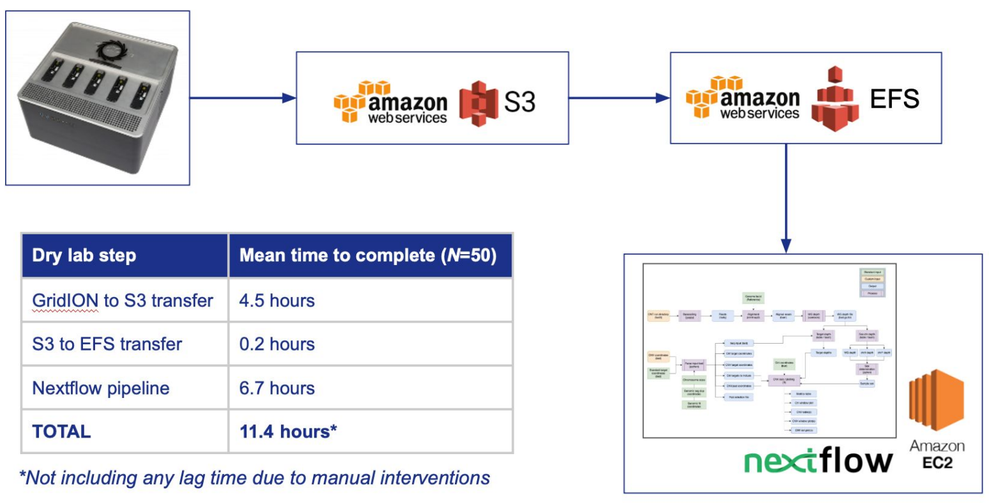
- Genomic copy number variants (CNVs) underpin many neurocognitive disorders, including intellectual disability, developmental delay, and autism spectrum disorders.
- Long read sequencing with Oxford Nanopore is a promising technology for clinical testing to identify causal variants in disease in a single laboratory assay, with the ability to test multiple genomic features, including sequence variants and methylation, and address a variety of existing challenges, such as read alignment in difficult-to-analyze genomic regions.

METHODS

- Sequenced 24 Coriell samples and 5 blood samples with known microdeletions/duplications using an ONT GridION sequencer with adaptive sampling.
- Each sample harbored at least one known CNV detected previously by microarray and short read WGS analysis; assayed a total of 35 unique CNVs (40kb to 155Mb in size).
- Robustness tested by assessing intra-run and inter-run concordance of samples.
- Developed a bioinformatic pipeline, implemented in Nextflow and executed on AWS, to assess the presence or absence of suspected CNVs using normalized read depth.
- A duplication or deletion was confirmed if the read depth across the suspected variant region was at least three standard deviations above or below, respectively, the mean depth of five genomic control regions.

Targeted ONT long read sequencing can reliably confirm copy number variants in a clinical test.

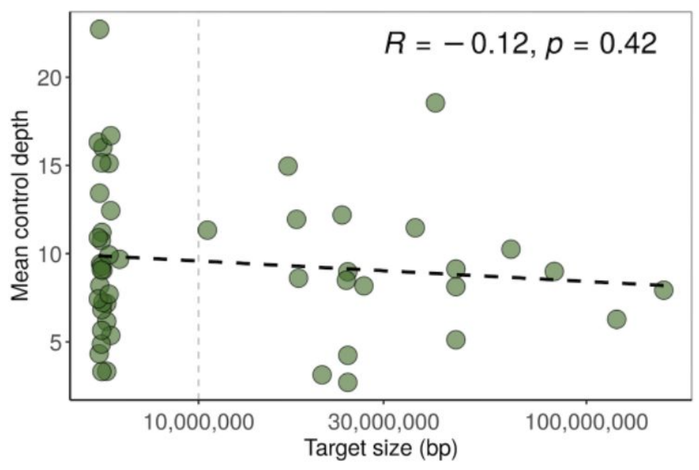
Figure 1. Schematic and timing of development and validation workflow



Dry lab step	Mean time to complete (*N*=50)
GridION to S3 transfer	4.5 hours
S3 to EFS transfer	0.2 hours
Nextflow pipeline	6.7 hours
TOTAL	**11.4 hours***

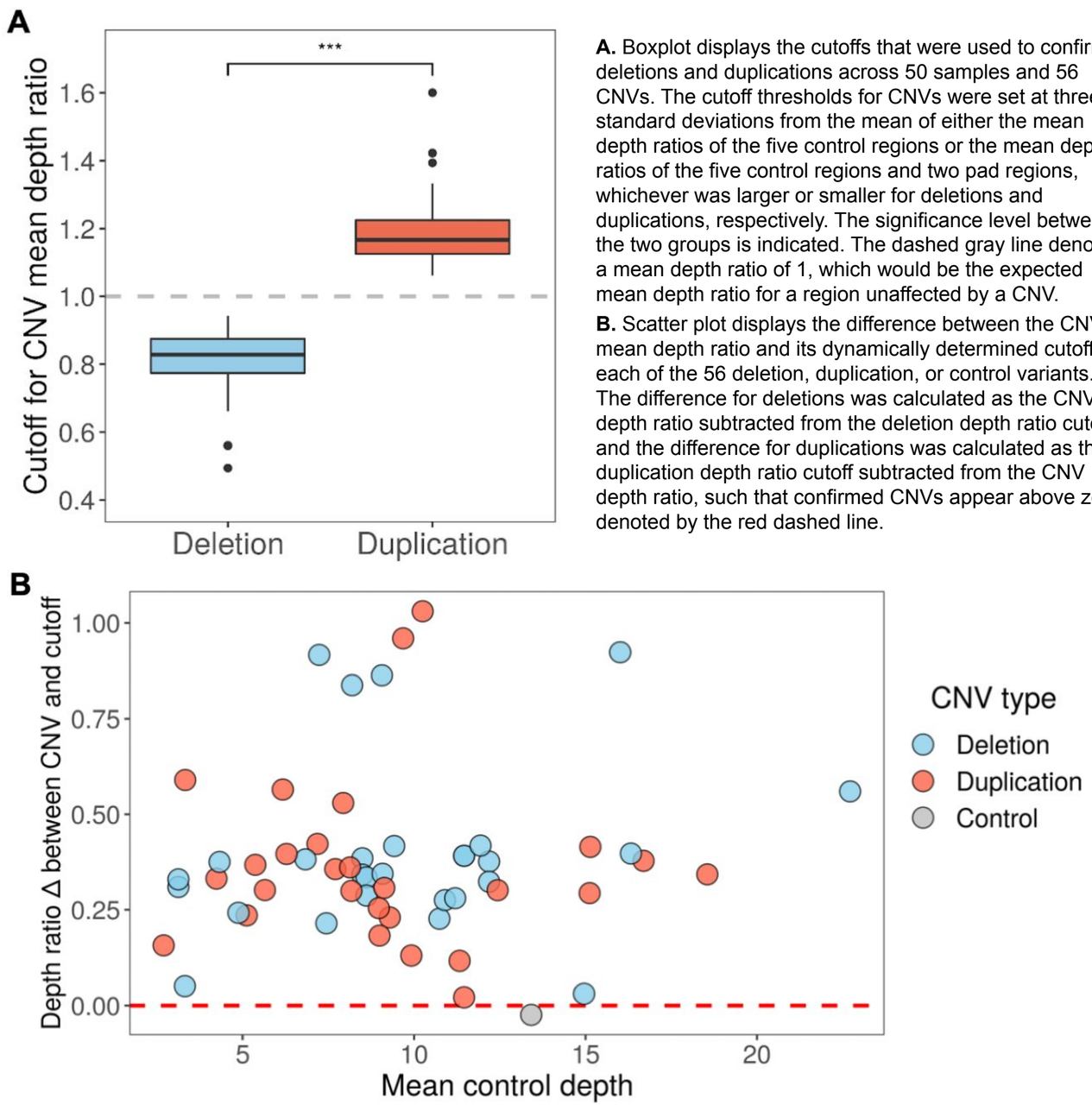
**Not including any lag time due to manual interventions*

Figure 2. Impact of total target size on on-target depth in control regions



Scatter plot displays the relationship between the total size targeted with adaptive sampling and the mean depth of the targeted control regions across all 50 samples. The linear regression trendline is plotted, and the Pearson correlation coefficient and its associated *p*-value is indicated on the plot.

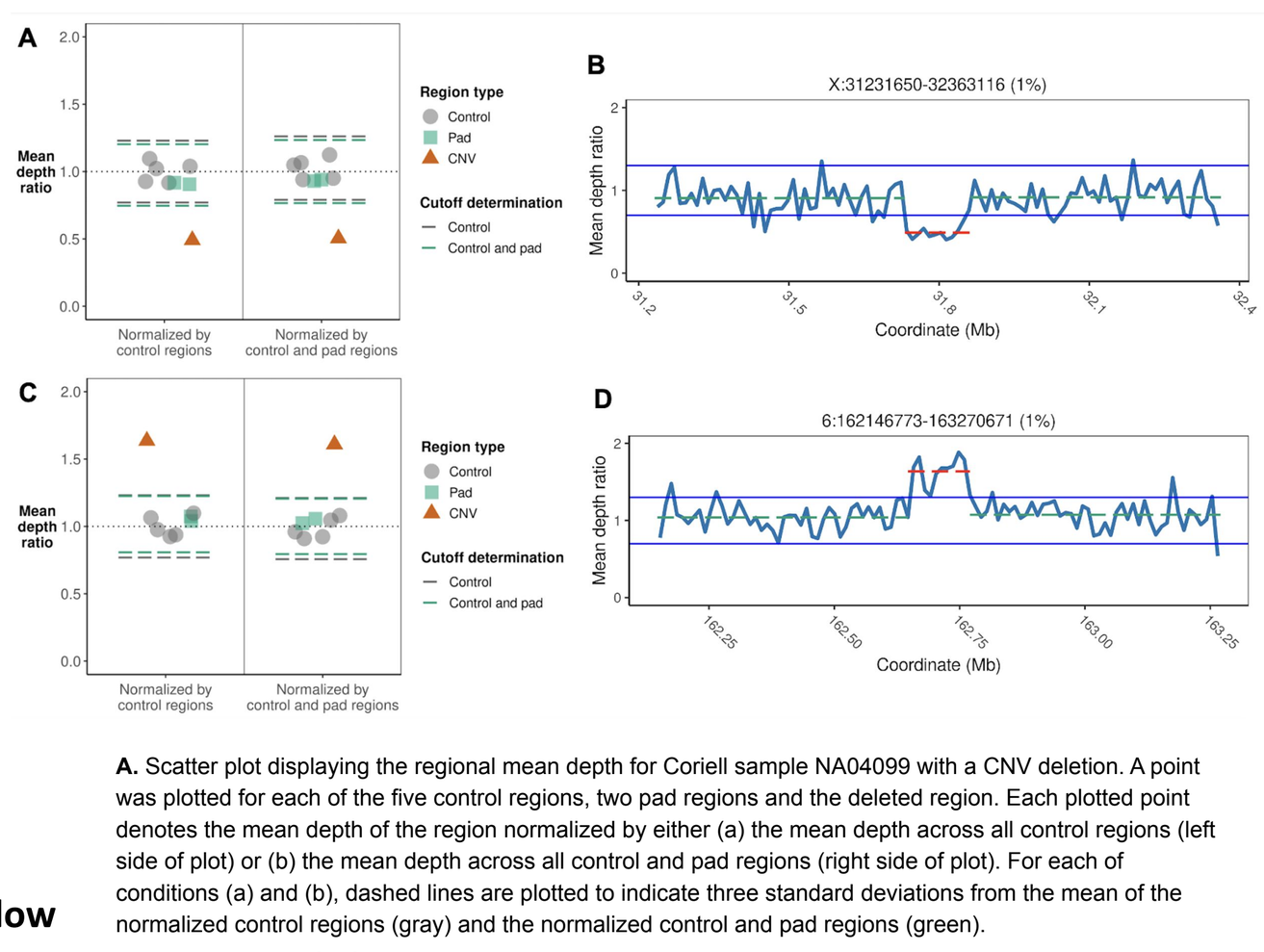
Figure 3. Cutoff thresholds for CNV confirmation



A. Boxplot displays the cutoffs that were used to confirm deletions and duplications across 50 samples and 56 CNVs. The cutoff thresholds for CNVs were set at three standard deviations from the mean of either the mean depth ratios of the five control regions or the mean depth ratios of the five control regions and two pad regions, whichever was larger or smaller for deletions and duplications, respectively. The significance level between the two groups is indicated. The dashed gray line denotes a mean depth ratio of 1, which would be the expected mean depth ratio for a region unaffected by a CNV.

B. Scatter plot displays the difference between the CNV mean depth ratio and its dynamically determined cutoff for each of the 56 deletion, duplication, or control variants. The difference for deletions was calculated as the CNV depth ratio subtracted from the deletion depth ratio cutoff and the difference for duplications was calculated as the duplication depth ratio cutoff subtracted from the CNV depth ratio, such that confirmed CNVs appear above zero, denoted by the red dashed line.

Figure 4. Example of CNV results for deletion and duplication sample

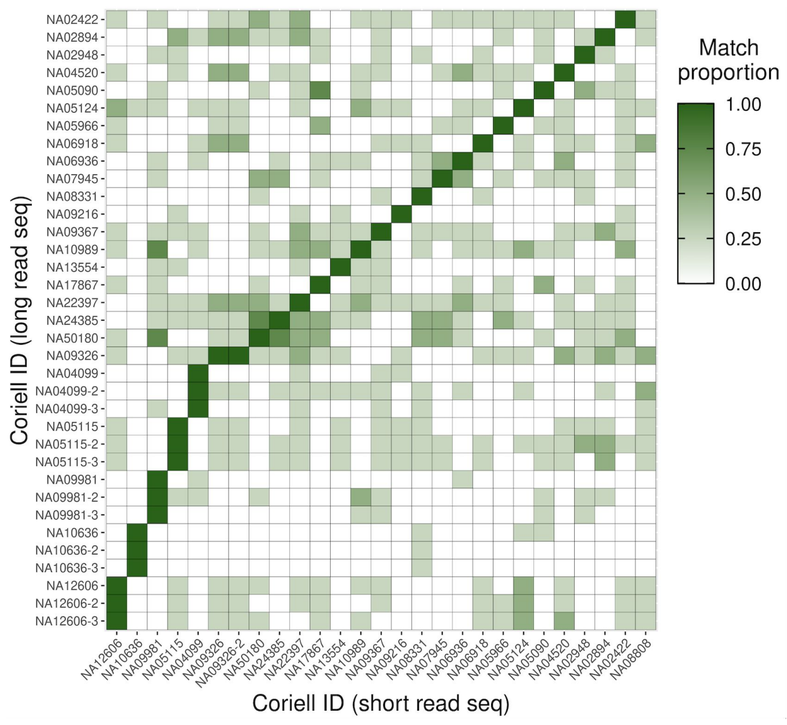


A. Scatter plot displaying the regional mean depth for Coriell sample NA04099 with a CNV deletion. A point was plotted for each of the five control regions, two pad regions and the deleted region. Each plotted point denotes the mean depth of the region normalized by either (a) the mean depth across all control regions (left side of plot) or (b) the mean depth across all control and pad regions (right side of plot). For each of conditions (a) and (b), dashed lines are plotted to indicate three standard deviations from the mean of the normalized control regions (gray) and the normalized control and pad regions (green).

B. Read depth plot for Coriell sample NA04099 displays the mean depth in the deleted region normalized by the mean depth across all control regions, calculated in non-overlapping windows that are 1% of the target size. The red dashed line indicates the mean depth ratio across the deleted region and the green dashed lines indicate the mean depth ratios of the pad regions directly adjacent to the deleted region.

C and **D.** Same as A and B, respectively, for a blood sample with a known duplication.

Figure 5. Preserving Sample ID



Heatmap shows the proportion of four control regions that match between Coriell samples sequenced previously on Illumina short read sequencing and in this study using ONT long read sequencing. The regions are matched based on SNV content. The analysis includes samples that were sequenced multiple times with long read sequencing and short read sequencing, as well as a sample that was sequenced using only short reads.

RESULTS

- Across 49 MinION flow cells, achieved an average on-target mean depth of 9.4X (range: 2.7X - 22.7X).
- Successfully confirmed all 55 of the expected CNVs across 49 samples (including replicates).
- Matched 26 unique Coriell samples using long read and short read data.

CONCLUSIONS/FUTURE DIRECTIONS:

- Results represent an initial clinical application of ONT long read sequencing to confirm CNVs.
- The cost of clinical long read sequencing can be reduced by multiplexing samples on higher-yield PromethION flow cells.
- Targeted ONT long read sequencing can enable improved variant discovery in difficult-to-analyze genomic regions, simultaneous methylation and sequence analysis, and ultimately an improved rate of clinical diagnosis.

Correspondence: Akash Kumar, akumar@myome.com