Advanta CFTR NGS Library Prep Assay Analytical Validation

Introduction

Cystic fibrosis (CF) is an autosomal recessive condition that results from variants in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Individuals with classical CF display symptoms that mostly affect the respiratory, digestive, and reproductive systems. Detection of CFTR variants in affected or carrier individuals has been performed with molecular testing methods such as allele-specific PCR and mass spectrometry. However, next-generation sequencing (NGS) for CFTR variant detection has been growing in adoption, leading to evaluations of greater variant diversity[1].

The Advanta™ CFTR NGS Library Prep Assay is designed to detect CFTR variants from genomic DNA samples using next-generation sequencing (NGS). Using the automated microfluidic capabilities of the Fluidigm Juno™ system, Advanta CFTR NGS libraries can be prepared in a highly efficient and scalable workflow. The assay provides targeted sequencing of the CFTR gene to cover 256 cystic fibrosis-causing single-nucleotide variants (SNVs) and insertions and deletions (indels)[2].

Analytical validation studies are critical for demonstrating robustness and reproducibility of assay performance in routine laboratory use. This paper describes the analytical validation of the Advanta CFTR NGS Library Prep Assay, including read metrics performance, variant call performance, and reproducibility. The study was performed by Q^2 Solutions® | EA Genomics (EA), a leading genomics research service provider. The study included a range of sample types, including genomic DNA, synthetic control DNA samples and DNA extracted from whole blood, saliva, and buccal swabs. Collectively, the samples directly tested 129 unique CF-causing variants. The results of this study demonstrate that the Advanta CFTR NGS Library Prep Assay is a robust and reproducible method for targeted sequencing of the CFTR gene from various sample sources.

Methods

Run Design

Independent runs were performed across several days with 15 Fluidigm LP 48.48 integrated fluidic circuits (IFCs) and three LP 192.24 IFCs. The runs were balanced across two operators, two reagent lots, eight library preparation days, three Illumina® MiSeq™ instruments, and eight flow cells. A subset of samples was also run at two sample inputs (see Sample Description and Preparation for details).

Sample Description and Preparation

Sample types used in analytical validation testing included purified genomic DNA (gDNA), synthetic DNA, and DNA extracted from whole blood, saliva, and buccal swabs. The following purified gDNA samples were obtained at the Coriell Institute for Medical Research (Coriell) from either the NIGMS Human Genetic Cell Repository, the NHGRI Sample Repository for Human Genetic Research, or the CDC Cell and DNA Repository: CD00003, CD00004, CD00005, CD00006, CD00008, CD00009, CD00010, CD00012, CD00013, HG00154, HG00376, HG00583, HG01092, HG01139, HG01620, HG01809, HG01853, HG02604, HG02661, HG02715, HG02882, HG03111, HG04141, HG04239, NA00130, NA00768, NA00897, NA00997, NA00998, NA00999, NA01012, NA01531, NA01707, NA01805, NA02828, NA04330, NA04346, NA06966, NA07228, NA07339, NA07381, NA07383, NA07441, NA07464, NA07469, NA07552, NA07732, NA07830, NA07857, NA07860, NA08342, NA11274, NA11275, NA11277, NA11278, NA11279, NA11280, NA11281, NA11282, NA11283, NA11284, NA11285, NA11286, NA11287, NA11288, NA11290, NA11370, NA11472, NA11496, NA11723, NA11761, NA11859, NA11860, NA12444, NA12585, NA12785, NA12889, NA12926, NA12960, NA12961, NA13423, NA13591, NA18668, NA18799, NA18800, NA18801, NA18802, NA18803, NA18886, NA18969, NA19092, NA19116, NA19130, NA19384, NA19448, NA19908, NA20513, NA20585, NA20737, NA20741, NA20745, NA20752, NA20836, NA20837, NA20905, NA20915, NA20924, NA20925, NA21069, NA21080, NA22162, NA23254, NA12878, NA24143, NA24149. Samples NA12878, NA24143, and NA24149 served as well-characterized samples that did not contain any CFTR variants. They are among the reference materials created by the Genome in a Bottle Consortium (GIAB) hosted by National Institute of Standards and Technology (NIST)[3]. 112 other gDNA samples from Coriell contained CF-causing variants but did not contain reliable information on positions negative for variants. All Coriell genomic DNA samples were run at 100 ng. Synthetic DNA plasmids (G211A, G211B, G211C, G211D, G211E, G211F) from Maine Molecular Quality Controls, Inc. (MMQCI) served to provide a greater diversity of CFTR variants to test, but some regions on MMQCI samples prevented proper variant detection (see analysis filters). MMQCI samples were run at a sample input of 30,000 copies. Finally, DNA extracted from whole blood, saliva, and buccal swabs reflect the sample types of the target user. Each sample extract was run at two DNA input quantities, 100 ng (n=117) and 60 ng (n=56). Like the CFTR variant-containing gDNA samples, reliable information on positions negative for variants in these samples was not available. All gDNA samples in the analytical validation that was run met the purity requirement of A260/280 ≥1.5. A no template control (NTC), which was a buffer, was included for each IFC to assess cross-contamination. A unique

barcode was added to each sample as an identifier during sequencing. Sample type breakdown of the 1,296 samples used for analytical validation is shown in Table 1.

Table 1. Sample type breakdown for analytical validation

Sample Type	Unique Samples	Total Samples
Coriell NIST reference samples (no CFTR variants)	3	132
Coriell CFTR variant-containing samples	112	732
MMQCI synthetic samples	6	234
Buccal	5	65
Saliva	5	58
Whole blood	5	57
NTC (run in singlet for each IFC)	N/A	18
Total	**136**	**1,296**

Assay Preparation

73 CFTR primer pairs were organized into pools and provided in the eight tubes included with the Advanta CFTR NGS Library Prep Assay Kit. 44 primer pairs were designed to interrogate single-nucleotide variants (SNVs) and small insertions or deletions (indels <7 bp) while an additional 29 primer sets were designed to detect large indel events (indels over 200 bp). The stock assays were prepared as described in the Fluidigm protocols[4,5].

Advanta CFTR NGS Library Preparation with Juno and Sequencing

Targeted specific amplification of the CFTR gene was performed using the Fluidigm Advanta CFTR NGS Library Prep Assay on the LP 48.48 IFC or LP 192.24 IFC as detailed in the Fluidigm protocols[4,5]. User-prepared sample mixes, which contain barcodes, and CFTR assay mixes were dispensed into the assigned IFC inlets. The loaded IFC was then run on the Juno system, which automatically combined the sample and assay mixes and ran the amplification script. Barcoded amplicons from all samples were harvested, pooled, and purified prior to the addition of sequencing adapters via a second PCR reaction. Libraries were analyzed using the Agilent® TapeStation, quantified, and normalized to 2 nM. Samples were sequenced across eight flow cells using three MiSeq instruments. Each flow cell had between 139 and 188 samples. All sequencing runs used MiSeq Reagent Kit v2 (Illumina) and adopted a 2 x 150 bp paired-end strategy. Mean amplicon read depths for all sample types are shown in Table 2.

Table 2. Mean amplicon read depth by sample type

Sample Type	Mean Amplicon Read Depth
Coriell genomic samples	2,576
MMQCI synthetic samples	1,469
Buccal	1,155
Saliva	1,389
Whole blood	3,105

Data Analysis Pipeline

Samples were demultiplexed and the resulting sample FASTQ files were trimmed to remove both low-quality reads and reads aligning to adapter sequences. Reads that met any of the following criteria were deemed low-quality and removed from the FASTQ: less than 25 bp in length post trimming, had a mean Q score below Q25, had greater than 25% of the read aligning to adapter sequence, or 95% of the read was identified as a homopolymer. Trimmed FASTQ files were aligned to the hg19 human reference and locally realigned. Variant calling was performed on the realigned BAM files. The aligned sequences went through primer trimming and amplicon depth calculations. The data was summarized using a custom R tool created by EA, and the amplicon coverage was calculated on the primer-trimmed sequences. Assay accuracy and concordance calculations were performed for the 44 assays designed to interrogate SNVs and indels. Indels in repetitive regions have multiple possible alignments, and the bioinformatics tools used in this validation report the leftmost alignment in a repetitive region. For a complex indel, the resultant call may be broken down into multiple smaller SNV/indel calls because this combination of SNV/indel optimizes the alignment score. Indels that met this criterion were left-aligned by moving the start position of the indel to the lowest possible coordinate value, and the co-ordinates were adjusted if the indels were located in repetitive regions or were a complex event.

The Fluidigm pipeline is described in the Advanta CFTR protocols[4,5] and analytical validation study guidelines[6]. Analysis of the sequencing data generated by EA using the Fluidigm pipeline produced results similar to the results that EA produced with its pipeline and that are described in this analytical validation report.

Analysis Filters

Analyses required that filtering be performed at the level of samples and variant calls before metrics could be calculated. Samples with very few reads were excluded from analysis. Using an EA demultiplexing tool, 14 samples (1.1% of 1,296 samples) were excluded from analysis for not meeting the minimum reads threshold. The threshold was defined by EA as 0.5% of the largest barcode read count.

An additional seven samples, which were analyzed for read performance metrics, were excluded from variant call metrics performance due to operator error during the library

preparation process. This error was determined by analyzing variant profiles of adjacent wells during the library normalization process. The mixed variant profile of samples prevented proper calculation of variant call metrics.

Samples were further filtered for variant calls at the level of base positions. Any base positions with read depth below 50x along with positions called as variants with a filter other than PASS were considered a no call (could not be confidently identified as either homozygous reference or a variant) and were excluded from any subsequent analyses. Positive variant calls also required a minimum allele frequency of 10%.

Some regions were blacklisted for MMQCI synthetic samples due to design incompatibility between the synthetic sample and select primers, which prevented the primers from properly annealing to the target. As a result, some variants were not detected in the synthetic samples.

Results

Read Performance Metrics

Read performance metrics included assay pass rate, amplicon uniformity, reads mapped to genome, reads mapped to target, and the NTC reads. Only the 44 amplicons designed for SNV and small indel detection were used to calculate these metrics. All sample types were used in the calculation of read performance metrics.

Table 3. Summary of read performance metrics

Attribute	Observed Performance
Assay pass rate	99.3% (gDNA) 97.8% (MMQCI)
Amplicon uniformity	99.4% (gDNA) 98.2% (MMQCI)
Reads mapped to genome	≥99.9% (all sample types)
Reads mapped to target	≥99.99% (all sample types)
No template control (NTC) reads	NTC undetectable • 15/18 IFCs (undetected) • 3/18 IFCs (<0.02% total seq reads)

Assay Pass Rate

The assay pass rate was calculated as the percentage of the targeted CFTR region that was covered at a read depth of at least 50x. It was expected that at least 95% of the target region would be covered at sufficient depth. 99.3% (1,026/1,033) of all genomic DNA samples (Coriell, buccal, saliva, whole blood) met the assay pass rate requirement. Additionally, 97.8% (224/229) of all MMQCI samples achieved sufficient targeted region depth. Overall read coverage was sufficient to confidently call SNV and indel variants in the targeted CFTR region across all sample types.

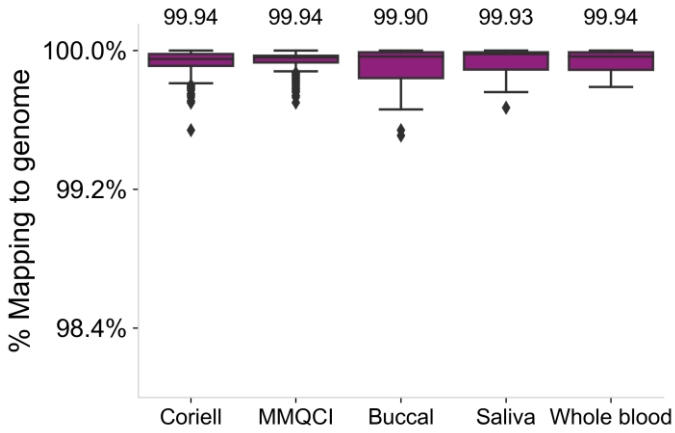
For this metric, Fluidigm reports the proportion of assays with at least 95% of the target region covered by >50x read depth.

Amplicon Uniformity

Amplicon uniformity was determined by counting the proportion of amplicons within 0.2x to 5x of the mean amplicon depth within the sample. 99.27% of Coriell samples, 98.2% of MMQCI samples, 99.48% of saliva samples, 99.24% for buccal swab samples, and 99.47% of whole blood samples met this criterion.

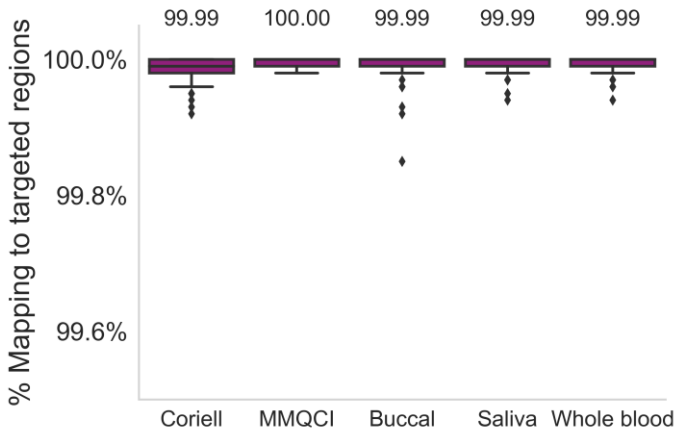
Reads Mapped to Genome

Mapping statistics were generated with an EA tool that aligns trimmed FASTQs to the human genome reference (hg19) and creates BAM files. Mapping to the genome was high, with an average mapping to genome of hg19 of 99.94% for Coriell samples, 99.94% for MMQCI samples, 99.90% for buccal samples, 99.93% for saliva samples, and 99.94% for whole blood samples.



Reads Mapped to Target

The number of reads that mapped to the CFTR target region was determined by calculating the proportion of reads mapped to the target region out of the total reads per sample. On-target mapping for all samples was very high, with an average on-target performance of 99.99% for Coriell samples, 100% for MMQCI samples, 99.99% for buccal swabs, 99.99% for saliva samples, and 99.99% for whole blood samples. This indicates that primers are specific and only amplify the region of interest.



No Template Control (NTC) Reads

The NTC read metric was the percentage of total reads assigned to the NTC divided by the total reads of the run. Each IFC included a NTC, which resulted in a total of 18 NTCs for this validation. 15 NTCs were never detected at the FASTQ demultiplexing step. Three NTCs were detected but comprised <0.02% of the total sequencing reads. This demonstrates that potential cross-contamination for this assay is consistently at a minimal level.

Alternatively, NTC reads can be calculated by using the ratio of target mapped reads of NTCs to the average of total mapped reads of DNA samples. Fluidigm adopts use of targeted mapped reads for calculation of the NTC read metric.

Variant Call Performance Metrics

Calculation of variant call metrics depended on the availability of truth data for which assay calls could be compared. Therefore, for sensitivity, sample types were limited to Coriell data and MMQCI samples. Additionally, only well-characterized Coriell NIST reference samples (NA12878, NA24143, and NA24149) and MMQCI samples, which both contained high confidence in known negative positions, could be used for calculation of specificity and accuracy. Reproducibility does not depend on the availability of truth data. Therefore, all sample types were used for calculation of reproducibility. This section pertains only to performance metrics for SNVs and small indels. Large indel assessment is discussed separately.

Table 4. Sample types used for calculation of variant call metrics

Sample Type	Sensitivity	Specificity	Accuracy	Reproducibility
*Coriell NIST reference samples	X	X	X	X
Coriell CFTR variant-containing samples	X			X
*MMQCI synthetic samples	X	X	X	X
Buccal, saliva, whole blood				X

* Samples with high confidence in known negative data

EA measured variant call performance for all variants covered by the Advanta CFTR NGS Library Prep Assay amplicons. However, Fluidigm recommends that variant call analysis be limited to CF-causing variants as specified in the CFTR2 database[2].

Sensitivity

Sensitivity was performed with Coriell and MMQCI samples. For Coriell samples without truth data available in the NIST database, truth data was taken directly from the Coriell website[7] and the 1000 Genomes Project (phase 3)[8]. Samples HG01853 and NA12878 had no variants reported for the CFTR gene in any database and were excluded from this analysis. Sample NA00130 had conflicting information for SNV positions between the Coriell website and the 1000 Genomes Project and was excluded from the analysis. Truth data for the MMQCI samples was obtained from the vendor. Sensitivity was defined as:

$$\frac{(TP)}{(TP + FN)}$$

Coriell samples had an overall sensitivity of 99.12% for SNVs and 99.71% for indels. The MMQCI samples had an overall sensitivity of 99.92% for SNVs and 96.13% for indels.

Specificity

Specificity was calculated using the three Coriell NIST reference samples (NA12878, NA24143, and NA24149) and MMQCI samples with confirmed truth for positions in the CFTR gene. Truth data for the Coriell samples was determined from the NIST data of high-confidence SNVs, small indels, and homozygous reference calls. Specificity was defined as:

$$\frac{(TN)}{(TN + FP)}$$

The three aforementioned Coriell samples had an overall specificity of 100% for SNVs and 100% for indels. The MMQCI samples had an overall specificity of 100% for SNVs and 100% for indels.

Accuracy

Accuracy was performed using three Coriell NIST reference samples (NA12878, NA24143, and NA24149) and the synthetic MMQCI samples. Accuracy was defined as:

$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$

The targeted result for accuracy was ≥99% for SNVs and indels of up to 6 bases in length and ≥90% for indels >6 bases in length. Coriell samples NA12878, NA24143, and NA24149 had an accuracy of 100% for SNVs and an accuracy of 100% for indels. MMQCI samples had an accuracy of 99.99% for SNVs and an accuracy of 99.99% for indels.

Reproducibility

Reproducibility assessed the reliability of the assay to return the same variant call results across several variables. This was calculated by performing pairwise comparisons across the study design variables: operator, reagent lot, library preparation day, IFC type, and input amount. Concordance was calculated only at positions with a coverage of at least 50x in both replicates being compared. If a position was marked as no call (could not be confidently identified as either variant or homozygous reference; see Analysis Filters) in either of the replicates being compared, the position was not included in the concordance calculation. As mentioned above, some regions were blacklisted for MMQCI synthetic samples due to design incompatibility between the synthetic sample and select primers. Concordance was calculated for SNVs and indels with the following formula:

$$\frac{\sum(Matches\ across\ all\ comparisons)}{\sum(Matches\ and\ Mismatches\ across\ all\ comparisons)}$$

The following concordant results are across all variables. Coriell samples were 99.93% concordant for SNVs and 99.46% for indel detection. MMQCI samples showed 98.69% concordance for SNVs and 96.43% for indel detection. For samples extracted from buccal swabs, saliva samples, and whole blood samples, SNV concordance was 98.47%, 98.40%, and 100%, respectively. Indels were not detected in these samples.

Table 5. Summary of variant call metrics

Attribute	Observed Performance
Sensitivity	99.1% SNVs (Coriell) 99.7% indels (Coriell) 99.9% SNVs (MMQCI) 96.1% indels (MMQCI)
Specificity	100% for all samples (SNVs and indels)
Accuracy	100% (Coriell) 99.99% (MMQCI)

Attribute	Observed Performance		
Reproducibility	**Sample**	**SNVs**	**Indels**
	Coriell	99.9%	99.5%
	MMQCI	98.7%	96.4%
	Buccal	98.5%	N/A
	Saliva	98.4%	N/A
	Whole blood	100%	N/A

Performance of Large Indel Detection

The Advanta CFTR NGS Library Prep Assay contains 29 amplicons that are designed to capture large indel events. The bioinformatic strategy used for detection of large indel events may vary by laboratory or institute. Coriell sample NA18668 was expected to have a 21 kb heterozygous deletion event resulting in the loss of exons 2 and 3. Amplification of the CFTR047_Del primer set was expected in samples positive for the deletion, and this was identified across all replicates, resulting in a 100% sensitivity for Coriell samples with known large insertion/deletion events.

A heterozygous deletion event of exon 2 and 3 was expected for MMQCI sample G211A, which was detected across all replicates. MMQCI sample G211F has a heterozygous deletion event of exon 25 and 26, which was not detected due to the amplicon design of this assay relative to the design of the synthetic sample. The primers that would detect this event are designed to anneal in intronic regions that were deleted within the sample. Therefore, capturing this event was outside the scope of this assay due to incompatibility between the primer and synthetic sample designs.

Conclusion

Analytical validation results show strong performance of the Advanta CFTR NGS Library Prep Assay. Read performance was excellent, based on assay rate, amplicon uniformity, mapping to genome, and mapping to target metrics. Likewise, we observed optimal performance for variant call sensitivity, specificity, accuracy, and reproducibility. Importantly, these parameters were evaluated with a diversity of samples, including gDNA extracted from buccal, saliva, and whole blood, which represent the primary sample types tested by the application. Taken together, this analytical validation performed by an external service laboratory demonstrates that the Advanta CFTR NGS Library Prep Assay is a prime tool for the next phase of CFTR molecular testing—the application of next-generation sequencing for variant detection.

For users who desire to perform their own analytical validation, study guidelines (PN 101-8063) are available. Example data may also be requested.

References

1 Brennan, M.L. and Schrijver, I. "Cystic fibrosis: A Review of Associated Phenotypes, Use of Molecular Diagnostic Approaches, Genetic Characteristics, Progress, and Dilemmas." *Journal of Molecular Diagnostics* 18 (2016): 3–14.

2 The variants covered are derived from the Clinical and Functional Translation of CFTR (CFTR2), available at cftr2.org (CF-causing variants from CFTR2_8August2016.xlsx).

3 Zook, J.M. et al. "Extensive sequencing of seven human genomes to characterize benchmark reference materials." *Scientific Data* 3 (2016): 160025.

4 Advanta CFTR NGS Library Preparation on the LP 48.48 IFC with Juno (PN 101-6270).

5 Advanta CFTR NGS Library Preparation on the LP 192.24 IFC with Juno (PN 101-6212).

6 Advanta CFTR NGS Library Prep Assay Analytical Validation Study Guidelines (PN 101-8063).

7 www.coriell.org

8 www.internationalgenome.org/phase-3-structural-variant-dataset

CORPORATE HEADQUARTERS
7000 Shoreline Court, Suite 100
South San Francisco, CA 94080 USA
Toll-free: 866 359 4354 in the US and Canada
Fax: 650 871 7152
fluidigm.com

SALES
North America | +1 650 266 6170 | info-us@fluidigm.com
Europe/EMEA | +33 1 60 92 42 40 | info-europe@fluidigm.com
China (excluding Hong Kong) | +86 21 3255 8368 | info-china@fluidigm.com
Japan | +81 3 3662 2150 | info-japan@fluidigm.com
All other Asian countries | +1 650 266 6170 | info-asia@fluidigm.com
Latin America | +1 650 266 6170 | info-latinamerica@fluidigm.com