Application Note

Genomics

Agilent

Trusted Answers

Evaluating cfDNA NGS Performance with the Agilent Avida DNA Workflow and Seraseq ctDNA Mutation Mix v4

Authors

Margherita Corioni, PhD,
Xiaomu Chem, MS, Heng Wang, MS,
and Grace Zhao, PhD,
R&D Scientists,
Agilent Technologies, Inc.

Collaborators from LGC Clinical Genomics

Andrew Anfora,
Associate Director of Product Management

David Merriam,
Commercial Director

Yves Konigshofer,
Director of Technology Development

Introduction

Analyzing cell-free DNA (cfDNA) using next-generation sequencing (NGS) has significantly advanced our understanding of cancer biology and facilitated the development of minimally invasive molecular diagnostics. However, conventional NGS workflows are often complex and time-consuming. Each step, particularly purification, can lead to DNA loss, making them challenging for low-input applications.

The Agilent Avida DNA workflow, powered by a new hybridization technology, eliminates the need for amplification and purification during library preparation before hybridization-based enrichment. This simplified workflow reduces PCR-induced bias and shortens the hybridization time to one hour without compromising recovery. This is essential for low-input analysis, which demands high sensitivity and specificity. Moreover, minimizing PCR amplification-related bias is advantageous for more accurate copy number variation (CNV) detection.

Having a stable and comprehensive reference material with accurate variant content is critical for demonstrating the analytical performance of cfDNA NGS and for monitoring quality control throughout product development. In this application note, we use the Seraseq ctDNA Mutation Mix v4 reference material due to its comprehensive coverage of cancer-related variants, reduced background noise compared to previous ctDNA references, and close resemblance to real cfDNA in term of size distribution and ligation efficiency.[1]

The Avida technology captures both positive (+) and negative (–) strands of the targeted DNA and uses unique molecular identifiers (UMIs) on each molecule for error correction and accurate molecule counting. Using UMIs, we report

- DNA molecule recovery at different input amounts across three oncology panels (26 kb, 345 kb, and 2.7 Mb), each sequenced at different depths.
- Variant detection with the three panels, plus an in-depth detection evaluation with the medium-sized panel (345 kb), which balances relevant oncology research variants and affordable sequencing depth.

The sequencing metrics reported in this application note can be generalized to other Avida catalog or custom panels in terms of expected feasible variant allele frequency (VAF) detection and molecule recovery for a given input and sequencing depth.

Methods

Samples and sample preparation

Seraseq ctDNA Mutation Mix v4 at 0% (wild type), 0.5%, and 5% VAF were used directly from the manufacturer-provided stock (p/ns 0710-3101, 0710-3099, and 0710-3100, SeraCare). The complete list of variants included in this reference material can be found on the SeraCare webpage. The 0.2%, 1%, and 2% VAF samples were created by diluting the 0.5% and 5% VAF stocks into the 0% wild type. The concentration of the Seraseq ctDNA Mutation Mix v4 was confirmed and quantified using the Agilent 4200 TapeStation system (p/n G2991A) with the Agilent D1000 DNA ScreenTape (p/n 5067-5582) and corresponding reagents (p/n 5067-5583).

Various input amounts of the reference material were enriched using the three Avida catalog panels from Agilent:

- Avida DNA Focused Cancer panel (p/n 5280-0050)
- Avida DNA Expanded Cancer panel (p/n 5280-0047)
- Avida DNA Discovery Cancer panel (p/n 5280-0044)

Each condition was tested in at least two technical replicates.

Panel details

- Avida DNA Expanded Cancer panel: Covers 105 relevant cancer genes and overlaps with 60 single-nucleotide variants (SNVs) and insertions/deletions (indels), seven translocations, and seven CNVs of the Seraseq ctDNA v4 reference material. This panel was preferred in this investigation due to its convenient size and overlap with the reference material.
- Avida DNA Focused Cancer panel: The smallest cancer panel (26 kb), optimized for high recovery of a focused set of cancer targets, mainly hotspots and exons, from 14 key genes in oncology.
- Avida DNA Discovery Cancer panel: The largest of the Avida cancer panels (2.7 Mb), covering over 680 key genes and biomarkers for discovery and assessment in oncology research.

The Focused and Discovery panels overlap with 23 and 66 SNVs and indels of the Seraseq ctDNA v4 reference material, respectively. Examples of allele frequency (AF) detection with these panel sizes will be demonstrated.

Library preparation and target enrichment

The ctDNA references were processed following the manufacturer's protocol for the Avida DNA kit (G9418A, Agilent) using the Focused, Expanded, or Discovery Cancer panels. Index PCR products were quantified using the D1000 ScreenTape before pooling for sequencing.

Data analysis

The Avida adapter contains UMIs, which are short sequences of random oligonucleotides used to tag each molecule in a sample library before PCR amplification. UMI barcode-based sequencing alignment was used to reduce the rate of false positive variant calls and improve the detection of true mutations present in the initial DNA input. This method uses PCR duplicates to group the sequencing reads by unique UMIs and determines an accurate consensus sequence considering the frequency and the quality score of the reads. The process makes use of duplication for sequence error correction to improve the quality of the data. The relationship between sequencing budget, duplication rates and variant detection is discussed in the results.

The sequencing reads were deduplicated using start/stop positional information and inline UMIs in single and/or duplex mode and aligned to the genome with bwa-mem.

The SNV and indels in the Seraseq ctDNA Mutation Mix v4 reference material were identified using VarDict (version 1.5.0). CNVs were analyzed with CNVkit (version 0.9.8), and fusions were called using GeneFuse (version 0.6).

In this study, detection rate (sensitivity) is defined as the successful variant call in the Seraseq ctDNA Mutation Mix v4 reference material covered by the analysis panel. The false positive rate (specificity) is defined as the variant call from the wild type (0%) reference.

Results

Selection of the number of reads for sequencing and UMI-based deduplication for error correction

This investigation used 10 to 20 ng of Seraseq ctDNA Mutation Mix v4 reference material in the Avida DNA workflow with the following Avida Cancer catalog panels: Focused (26 kb), Expanded (345 kb), and Discovery (2.7 Mb). The amount of sequencing was adjusted based on panel size and DNA input amount. Table 1 shows an example of the correlation between the number of sequencing read pairs and the design content average raw median coverage in the study.

The sequencing read pairs budget is an essential parameter for variant detection analysis that must be appropriately selected to ensure sufficient coverage for a given variant detection cutoff. Typically, higher input DNA amounts and lower AF detection require a higher read budget. As described in the method section, the number of sequencing reads also directly correlates with the accuracy of the UMI deduplication process, which selects the most accurate consensus sequence for each unique UMI recovered.

Considering the input amount and the panel size at the capture step, sample sequencing depth should be selected so that the percentage of duplicates at deduplication approaches or exceeds 80% (for UMI single-mode deduplication). This allows for exhaustive screening of unique UMIs in the sample and accurate UMI unique-consensus assignment at deduplication.

For detection of low frequency variants using the power of inline UMI deduplication accuracy, it was found that with duplication percentages significantly lower than 80% a substantial fraction of the molecules is not sequenced and therefore does not contribute to variant calling. Under-sequencing the sample results in only partial visibility into the complexity of the library (the number of unique molecules in the sample), making it more likely that rare variants are not detected. Increasing the sequencing depth to significantly exceed 80% duplication offers limited benefits to the recovery of more unique molecules and, as such, in understanding the library composition. But, if feasible, additional sequencing is not discouraged, as it can help increase the size of each UMI family, supporting error correction.

Table 1. Impact of sequencing depth on design content median coverage and duplication rates for unique molecules consensus correction (by UMI single-mode deduplication) with the Agilent Avida cancer panels. Sequencing budget is expressed in read pairs. Total reads budget is twice the read pairs amount.

Panel	Size [kb]	Input [ng]	Sequencing Read Pairs [M]	Raw Median Coverage [k]	Duplication [%]
Avida DNA Focused Cancer	26	20	15	40	83
		20	10	27	76
		20	5	13	60
Avida DNA Expanded Cancer	345	10	50	14	88
		10	40	10.6	78
		10	20	5.3	62
	345	20	80	21.6	81
		20	50	14	72
		20	40	10.8	66
		20	20	5.4	47
Avida DNA Discovery Cancer	2700	20	400	19	80
		20	200	9	64
		20	100	4.7	45

General performance of three cancer panels with 20 ng reference input

Capturing the Seraseq ctDNA Mutation Mix v4 reference libraries with the Avida DNA Expanded Cancer panel using a 20 ng input and 80 M sequencing read pairs (160 M total reads) budget results in a mean coverage of 4000x across the panel (Table 2). Along with high coverage depth, the coverage is also uniform with more than 95% of the targeted bases having deduplicated coverage that is greater or equal to half the design mean. This means that over 95% of the region across the panel, including both exonic and intronic regions, have coverage exceeding 2000x.

Consequently, for a 0.2% AF with a 20 ng input, the average number of copies at any locus is expected to be at least four copies for 95% of the regions, and eight copies of the variant on average. These factors enabled a detection rate of 94% for a 0.2% AF at 20 ng input when analyzing all 60 SNV and indels in the Seraseq ctDNA Mutation Mix v4 reference covered by the Expanded panel (Table 2).

Using 20 ng of 0.2% VAF reference DNA as input (covering 23 variants) for the Avida DNA Focused panel (26 kb) with a 10 M read pairs budget (20 M total reads), the detection rate for SNV was 100% (Table 2). With an average UMI coverage of 6000x (Table 2), 12 copies of variants are expected, making the 100% detection rate consistent with expectations.

For the large Avida DNA Discovery panel (2.7 Mb), using 20 ng of input at 0.5% VAF and sequenced with approximately 170 M read pairs (340 M total reads), the detection rate for SNVs and indels (covering 66 variants) was 94% (Table 2). Following the same logic as the Expanded and Focused panel detection, given the input, panel size, and sequencing budget, an average UMI coverage of 2600x was achieved, with an expected 13 copies of variants on average. It is noted that the duplication rate with the given sequencing depth for this relatively large panel is only 68% (Table 2). Deeper sequencing will increase UMI recovery and likely enhance sensitivity.

The wild type (0%) reference was analyzed with the same input and similar sequencing depth and showed low false positive detection. One sample had a single false positive while a second had four false positives calls using 20 ng and the Focused and Expanded panels at 10 M and 80 M read pairs, respectively (equivalent to 20 M and 160 M total reads). No false positives were detected with the other combinations of panel, input, and sequencing depth.

Table 2. General performance metrics for the Agilent Avida DNA Focused Cancer, Avida DNA Expanded Cancer, and Avida DNA Discovery Cancer panels using 20 ng of Seraseq ctDNA Mutation Mix v4 reference input material: efficiency of molecular recovery (UMI-deduplicated coverage) and sensitivity of variant detection.

Panel	Size [kb]	Input [ng]	SNV/Indels Covered	Expected VAF [%]	Read Pairs [M]	Raw Reads Coverage {k]	UMI-Deduplicated Coverage [k]	Duplication [%]	Detected VAF [%]	Detection Sensitivity [%]
Avida DNA Focused Cancer	26	20	23	0.2	10	**27**	**6**	**78**	**0.2**	**100**
Avida DNA Expanded Cancer	345	20	60	0.2	80	**21**	4.3	81	0.2	94
Avida DNA Discovery Cancer	2700	20	66	0.5	172	**7.5**	**2.6**	**68**	**0.5**	**95**

The sequencing budget shown here (read pairs) was selected to allow adequate error correction and unique molecule recovery for proper sample interrogation (% duplicates). Sequencing budget is expressed in read pairs. Total reads budget is twice the read pairs amount.

Avida DNA Expanded Cancer panel detection with different combinations of input and sequencing budget

When looking for variants at low frequency, as in this case, it is strongly recommended to accurately determine the sequencing budget needed. This should be based on the combination of input amount, AF of interest, and the size of the capture panel. Ideally, the raw median coverage from the analysis should be high enough so that the percentage of sample duplicates is in the range of 80% or higher. This indicates that an exhaustive number of unique molecules has been screened and that the recovered unique molecules are identified with precision, so that the sample of interest can be interrogated with accuracy (see Table 2).

Acknowledging that the sequencing depth can be limited due to various factors, this application note provides examples of down-sampling the data with the respective duplication percentages and the impact on AF detection limits. This demonstrates the effect of the number of reads on variant detection rates and helps plan for the most accurate analysis experience possible (see Tables 1 and 3A).

Sometimes the DNA input into the assay can be limited, or the sequencing budget can be hard to estimate, all variables that impact the detection capability of the method. Tables 3A and 3B provide guidance of various combinations of inputs, sequencing budgets, and VAF detection rates observed with the Avida DNA Expanded Cancer panel.

Table 3.

A. Variant allele frequency (VAF) detection obtained with Avida DNA Expanded Cancer panel and Seraseq ctDNA Mutation MIx v4 reference input range of 20 to 40 ng.

Input [ng]	Sequencing Read Pairs [M]	Duplication [%]	VAF [%]		
			0.1	0.2	0.5
40	160	82	92	98	
	50	26	86	96	
	25	13	73	91	
	13	6	51	76	
20	80	81		94	100
	50	51		92	100
	25	25		87	98
	13	13		69	96

Impact of the sequencing real budget on error correction and number of molecules recovered (indicated as % duplicates), and variant detection sensitivity with VAF in the low range of 0.1 to 0.5%.

B. Variant allele frequency (VAF) detection obtained with Avida DNA Expanded Cancer panel and Seraseq ctDNA Mutation Mix v4 reference of ultra-low input in the range of 1 to 10 ng.

Input [ng]	Sequencing Read Pairs [M]	Duplication [%]	VAF [%]			
			0.5	1	2	5
10	46	81	96			
	32	80		98		
	39	78			99	
5	30	86	82			
	57	93		97		
	68	93				100
3	39	92		98		
1	33	92			98	
	11	94				93

Examples of sequencing read budget driving molecule recovery and error correction (indicated by % duplicates), and sensivitity of detection with VAF in the range of 0.5 to 5%. The Seraseq ctDNA Mutation Mix v4 reference covered by the Expanded panel includes 7 CNVs and 7 translocations. The CNV detection rate for both inputs was 100% (Table 3C). The average fold change for CNVs was around 1.1-fold for 0.2% and 1.3-fold for 0.5%. For translocation detection, the rates were 93% and 79%, respectively. No false positives were detected from the wild type (0%) reference DNA with comparable input and sequencing depth.

C. CNV and translocations.

Variant Type	VAF [%]	Input [ng]	Sequencing Read Pairs [M]	Duplication [%]	Detection Sensitivity [%]
CNV	0.2	20	80	90	100
CNV	0.5	10	66.6	81	100
Translocation	0.2	20	80	90	93
Translocation	0.5	10	66.6	81	79

Target coverage statistics for the Avida DNA Expanded Cancer panel

To provide a more general overview of the enrichment characteristics of the reference input material using the Avida DNA Expanded Cancer panel, additional statistics of coverage by target are presented. Figure 1 reports the coverage data for the enrichment target regions using a 20 ng reference input with the Expanded panel and 80 M read pairs (160 M total reads) allocated at sequencing. 100% of the regions targeted had 1000x coverage, 97% had 2000x coverage, and 87% had more than 3000x coverage. The average coverage across the entire panel was 4425x. More than 87% of the regions had more than 3000x coverage (Figure 1A).

Figure 1B illustrates the average coverage of key cancer-related genes using the Expanded panel. The overall exonic content regions had a coverage of at least 3000x on average.

Figure 1C depicts the detection of select variants in the reference material overlapping with the Expanded panel and the genes shown in Figure 1B. The input material is 20 ng with 80 M read pairs (160 M total reads) allocated at sequencing, and the expected allele frequency (AF) tested was 0.5% and 0.2%.

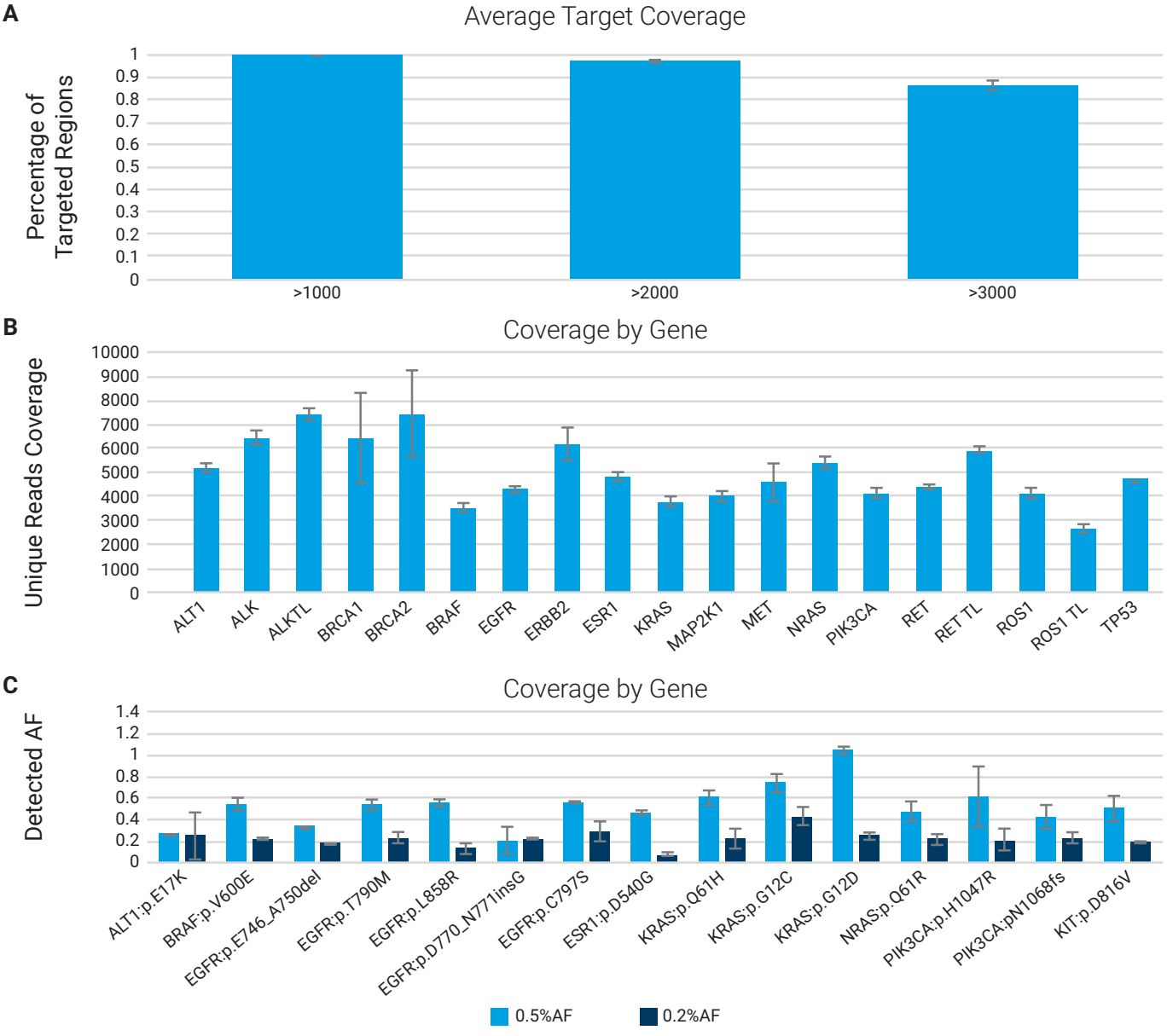

Figure 1. Recovery of selected genes from 20 ng input of Seraseq ctDNA Mutation Mix v4 reference material enriched by the Agilent Avida DNA Expanded Cancer panel. A. Median coverage statistics by target region (intended as unique reads coverage following UMI-deduplication) calculated from 80 M sequencing read pairs (160 M total reads). The median coverage is N of 6 replicates. **B.** Unique reads coverage aggregated by gene for key signature cancer-related genes, calculated from 80 M sequencing read pairs (160 M total reads), ("TL" indicates non-exonic targeted content intentionally included to boost translocation detection). **C.** Detection of variants in the reference versus the expected at 0.5% and 0.2% AF from 80 M sequencing read pairs (160 M total reads). The mean AF condition is N of 2 replicates from 20 ng of input material.

Conclusions

Evaluating the analytical performance of a cfDNA NGS method is a critical step and using a high-quality reference material is essential for this process. Seraseq ctDNA Mutation Mix v4 has proven to be an efficient and comprehensive tool for assessing analytical performance in variant detection with the Avida DNA workflow, due to its variety of variants and low background noise.

As demonstrated, the Avida DNA workflow can consistently recover sufficient unique DNA molecules to allow reliable variant calls at 0.5% and 0.2% VAF with a flexible input range of 10 to 20 ng of Seraseq ctDNA reference input.

It has been shown by down-sampling the sequencing data and performing variant detection with fewer than the recommended number of read pairs that the Avida technology still provides accurate and valuable measurements of the sample composition. This allows for some flexibility in the input amount and the sequencing budget allocation estimates, helping the interrogation of precious and limited samples.

As shown in this analysis, the Agilent Avida system easily tolerates a flexible range of input amounts without requiring protocol modifications or experiencing performance loss (see Table 3), making it an ideal solution for liquid biopsy applications and other investigations where single sample use and maximal efficiency of molecule recovery are important.

References

1. Ruminski, L.; et al. *Next generation liquid biopsy reference material performance across NGS assays and platforms.* Presented at AACR Annual Meeting, **2024**. Available for download at: https://digital.seracare.com/aacr-liquid-biopsy

www.agilent.com/chem/avidadna

PR7001-3473

Published in the USA, December 12, 2024
5994-7953EN