APPLICATION NOTE

CNV CALLING WITH LoopCap™ DNA TARGET CAPTURE KITS AND OPEN-SOURCE ANALYSIS TOOL CNVkit

A PRACTICAL GUIDE TO LEVERAGE LoopCap TECHNOLOGY FOR CNV ANALYSIS

Summary

Copy number variations (CNVs) play a vital role in genetic analysis, as they involve duplications or deletions of large segments in the genome. Ranging from hundreds to millions of base pairs, CNVs have significant implications in research and clinical settings. They have been linked to various disorders such as developmental abnormalities, neurodegenerative diseases, cancer, and pharmacogenomic responses. Detecting and characterizing CNVs is crucial for a comprehensive understanding of the genome.

In this application note, we provide researchers and clinicians with a practical guide to leveraging LoopCap technology, a MIP-based approach to targeted sequencing, for CNV analysis. Using open-source CNV calling tools, we demonstrate how LoopCap technology is a high-performance, high-throughput, and cost-effective sequencing choice for first-tier analysis for CNV detection.

molecular loop

Introduction

Array-based technologies have been used widely for CNV analysis since the late 1990s as they are affordable and relatively high-resolution assays for CNV detection. However, these technologies have technical limitations which have allowed for next-generation sequencing (NGS) technologies to replace them as first-line screening assays in recent years, with whole genome sequencing (WGS) as the primary strategy for NGS CNV detection.

When compared to WGS, targeted techniques offer lower cost, higher coverage, and less complex data analysis, making them ideal for many clinical applications. Unfortunately, traditional hybridization-capture sequencing techniques pose technical challenges for CNV calling. Most notably, random fragmentation used during sample preparation results in variation in read depth between samples, making CNV calling based on sequencing depth difficult. Molecular inversion probe (MIP) targeted sequencing uses highly specific targeting arms to generate library molecules containing only the region of interest, with no random fragmentation in the workflow. This results in more consistent relative read depth for the same probe across multiple samples than would be expected in a traditional targeted sequencing approach like hybridization-capture. The utility of MIP targeted sequencing for CNV calling has been demonstrated in several studies[5,6,7] and could enable a more economical and high-throughput approach to CNV detection when compared to WGS or traditional targeted sequencing techniques.

Although there are several open-source tools readily available for calling CNVs from WGS data, most of these tools have limitations with respect to CNV calling from targeted sequencing data.[1] As a result, many users are inclined to develop and optimize their own in-house tools for CNV calling.[2,3]

Here, we demonstrate how to leverage LoopCap technology for CNV analysis. We present an approach to CNV calling using LoopCap sequencing data and CNVkit, an open-source CNV caller. We highlight the essential steps in the analysis pipeline, from library preparation to data interpretation, and discuss the key considerations and challenges encountered during CNV detection.

Materials and Methods

The LoopCap DNA Target Capture Kit (ML4100) and Core Carrier Screening Panel were used to prepare target capture libraries from 150 ng of human genomic DNA. The capture targeted the 113 Tier 3 genes recommended by the American College of Medical Genetics and Genomics (ACMG) for carrier screening.* A total of 95 Coriell cell lines were included in the analysis, and among them, four samples had well-characterized CNVs (Table 1).

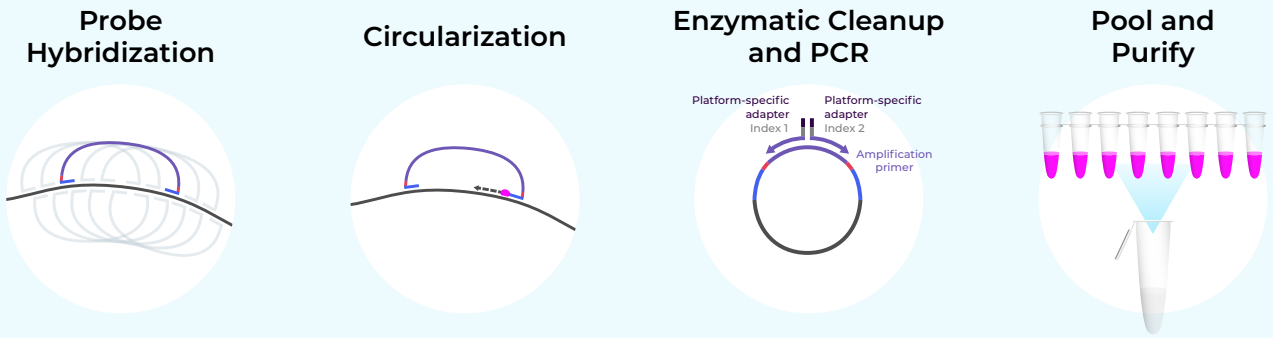

Figure 1. LoopCap DNA Target Capture workflow. Probes are hybridized to the target in a 4 – 16 hour hybridization step, followed by enzymatic circularization, digestion of non-circular DNA, and PCR to add platform-specific, unique dual-indexed barcodes and amplify the library. Next, samples are pooled and a SPRI bead purification is performed on the library pool, generating a sequencing-ready library.

Libraries were sequenced on the Illumina NovaSeq system (SP flow cells) with 2 x 150 bp read length. Data analysis was performed with a Picard/Samtools pipeline. Sample fastq files were normalized, deduplicated, and additional mapping quality filtering (map quality >10) was performed by Samtools. All 95 samples achieved mean coverage of >980X, with >94% of target covered at >100X.

CNV calls were made using CNVkit, configured using the best practices guidelines for amplicon-based libraries with additional bin tuning for target size and depth of coverage. A confidence threshold of 0.002 was used for CNVkit segmentation. Well-categorized CNVs from 16 experimental Coriell samples were detected against a control built using 79 categorized normal samples. Results were analyzed using Bedtools and custom VCF analysis software. Four of the experimental Coriell samples each contained one well-categorized CNV, each of various sizes (Table 1), and the other experimental controls contained either known structural variants or no known variants.

Results of our internal testing using LoopCap technology were compared to hybridization-based methods as reported in Gabrielaite et al,[1] and Gordeeva et al[4] (Table 2).

*Although SMN1/2 CNV calling has been previously demonstrated with MIP-based technologies elsewhere[8,9] and are included in the Molecular Loop Core Carrier Screening Panel, SMN1/2 were excluded from this analysis due to a lack of accurate ground truth reported from Coriell during model building.

Results

CNV calling performance was evaluated based on three key metrics: specificity (the method's ability to accurately identify regions without CNVs), recall (the method's capability to detect true CNVs), and precision (the accuracy of the called CNVs). These metrics provide insights into the overall performance and reliability of the CNV calling process.

For each of the 16 test samples, the 2285 target intervals were categorized as either True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN). To determine specificity (TN/[TN+FP]), categorized intervals were summed across samples. Precision (TP/[TP+FP]) and recall (TP/[TP+FN]) were calculated using contiguous calls and known positive CNV intervals.

Table 1. Coriell DNAs with known CNVs included in this study.

Sample	Chromosome	CNV	Result
NA21081	chr11	HBB 16kb_full_gene_del	Called
NA11661	chr17	GAA_delExon18	Not Called
NA02533	chr19	MCOLN1_6.4kb	Called
NA18668	chr7	CFTR_delExon2-3	Called

We achieved recall of 75% (3/4) and precision of 21% (3/14). Our study also showed excellent specificity (99.2%, 32644/36533), indicative of a low false positive rate and accurate identification of non-CNV regions.

While recall is high compared to other tools (see Table 2) and demonstrates a significant portion of CNVs being successfully detected, it does indicate the presence of a false negative in one sample, which CNVkit was unable to call (Table 1).

The observed precision of 21% is remarkably high considering the distribution of negative target (*n*=32644) to positive target (*n*=4) intervals in this study. In first-tier CNV analysis, results are commonly validated using array-based or MLPA approaches. For a screening assay followed by a confirmatory test, prioritizing sensitivity, even at the cost of precision, is crucial.

Discussion

In order to benchmark the performance of the MIP-based LoopCap method, we compared our results from our core carrier screening panel to published recall and precision data from several hybridization-based whole exome sequencing protocols (WES).[1,4] Each study compared the

performance of multiple CNV calling protocols and the results are summarized in Table 2. It should be noted that the capture target size in the LoopCap study (0.35 Mb capture target size) is significantly smaller than the WES data (>35 Mb capture target size) analyzed in these two publications. However, we were unable to find published data for CNV calling from similarly-sized hybridization-based capture panels.

In the first study,[1] recall values obtained were below 30% for all tools except CNVkit, which achieved a recall of 40%. Precision was similarly very poor for all CNV calling tools, with precision of less than 10% in all tools except CNVkit, which achieved precision of 39%. In the second study,[4] 3 of the tools showed recall values >25% (exomeCopy, EXCAVATOR2, and FishingCNV), but of those three, only EXCAVATOR2 had precision >5%.

In the context of clinical front-line testing, where positive screening results are followed by confirmatory testing using orthogonal methods, the importance of recall and precision becomes evident. In this scenario, achieving higher recall scores (with fewer false negatives) takes priority over higher precision scores (with a larger proportion of false positives). This approach ensures that potential cases are not missed (minimizing false negatives) and allows for further validation and confirmation of positive results.

Conclusion

CNV calling utilizing LoopCap and open-source tooling achieves excellent performance for detecting CNVs over known target regions, overcoming the recall limitations held by hybridization-based capture.

This makes LoopCap technology an ideal choice for first-tier analysis for CNV detection that allows for high-performance, high-throughput, and cost-effective sequencing compared to whole genome sequencing or hybridization-based capture.

Table 2. Comparison of CNV calling performance between LoopCap and hybrid capture protocols as reported in Gabrielaite et al[1] and Gordeeva et al.[4]

Source	Enrichment Method	CNV Analysis Tools	Recall	Precision
Internal testing	LoopCap	CNVkit	75%	21%
Gabrielaite et al.[1]	Agilent SureSelectXT Clinical Research Exome kit	CNVkit	40%	39%
		CLC	26%	2%
		GATK_gCNV	21%	0%
		Manta	10%	8%
		ExomeDepth	8%	1%
		cn.MOPS	6%	1%
		CODEX2	6%	1%
Gordeeva et al.[4]	Various hybrid-capture enrichment methods	exomeCopy	65%	5%
		EXCAVATOR2	48%	35%
		FishingCNV	27%	4%
		ExomeDepth	15%	59%
		CODEX	12%	58%
		cn.MOPS	7%	80%
		CNVkit	4%	68%

References

1. Gabrielaite, M. et al. A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. *Cancers (Basel)*. 2021 Dec 14;13(24):6283. doi: 10.3390/cancers13246283. PMID: 34944901; PMCID: PMC8699073.
2. Sorrentino, E. et al. CNV analysis in a diagnostic setting using target panel. *Eur Rev Med Pharmacol Sci.* 2021 Dec;25(1 Suppl):7 – 13. doi: 10.26355/eurrev_202112_27328. PMID: 34890029.
3. Singh, A.K. et al. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Med Genomics* 14, 214 (20 21). https://doi.org/10.1186/s12920-021-01059-x
4. Gordeeva, V. et al. Benchmarking germline CNV calling tools from exome sequencing data. *Sci Rep* 11, 14416 (2021). https://doi.org/10.1038/s41598-021-93878-2
5. Hitti-Malin, R. J. et al. Using single molecule Molecular Inversion Probes as a cost-effective, high-throughput sequencing approach to target all genes and loci associated with macular diseases. *Human Mutation*, 43, 2234 – 2250. https://doi.org/10.1002/humu.24489
6. Reurink, J. et al. Molecular Inversion Probe-Based Sequencing of USH2A Exons and Splice Sites as a Cost-Effective Screening Tool in USH2 and arRP Cases. *Int. J. Mol. Sci.* 2021, 22, 6419. https://doi.org/10.3390/ijms22126419
7. Neveling, K. et al. BRCA Testing by Single-Molecule Molecular Inversion Probes, *Clinical Chemistry*, Volume 63, Issue 2, 1 February 2017, Pages 503 – 512, https://doi.org/10.1373/clinchem.2016.263897
8. Boyden, E. et al. High-throughput screening for SMN1 copy number loss by next-generation sequencing; (#2597W). Presented at the Annual Meeting of The American Society of Human Genetics, October 22, 2013 in Boston, MA.
9. Boyden, E. et al. High-throughput screening for SMN1 “2+0” copy number status by next-generation sequencing; (#3240F). Presented at the Annual Meeting of The American Society of Human Genetics, October 20, 2016 in Vancouver, Canada.

For more information:
www.molecularloop.com | sales@molecularloop.com

molecularloop

Molecular Loop Biosciences, Inc.
300 Tradecenter Drive, Suite 5400, Woburn, MA 01801

sales@molecularloop.com
www.molecularloop.com