Effective and Accurate CNV Calling Using the PiVAT Bioinformatics Platform

Performance analysis of MET, MYC, and ERBB2 biomarkers

INTRODUCTION

Copy number variation (CNV) has been identified as a consequential type of structural variation in the human genome. These variations include additional copies (duplications) and losses (deletions) of genetic sequence ranging from 100 base pairs (bps) to 3 mega base pairs (Mbps) in length. Genetic diseases can arise from CNV-driven alterations in gene expression, demonstrating an urgent need for improved diagnostic methods for detection and therapeutic selection.

Previous studies demonstrated that copy number variation in genes such as MET, MYC, and ERBB2 can serve as biomarkers for cancer[1]. Gene amplification of these proto-oncogenes can lead to an activation of transcription factors and upregulation of protein products that drive abnormal cell proliferation. This leads to cancer progression and poor patient survival.

With the goal to help clinicians and doctors diagnose patients in the early stages of disease progression in a reliable, non-invasive, and systematic way, the Pillar variant analysis toolkit (PiVAT) provides a bioinformatics tool to detect CNVs.

PURPOSE OF EXPERIMENTS

Advances in next-generation sequencing (NGS) allow for more accurate detection of CNVs and for a deeper understanding of their relationship to tumor and disease susceptibility. Several metrics, including mapping rate, on-target rate, and coverage uniformity were analyzed to determine the target capture efficiency and target enrichment. Mapping rate refers to the percentage of reads that correctly map to the reference genome. The on-target rate refers to the percentage of mapped reads that align to targeted regions of interest. Coverage uniformity describes the variation in read-depth coverage across each base in the region of interest. We investigated these metrics to assess the reliability of CNV detection using our sequencing methods and PiVAT software.

PiVAT can detect duplications above 2.4n and deletions below 1.6n with the intermediate range (0.8 – 1.2 copy number ratio) considered to be at the normal inheritance level. It is important to

note that PiVAT is designed to not report CNVs unless the results show a unique significance in amplification or deletion of genes and gene segments.

The goals of this study are to 1) confirm panel stability over various DNA input amounts by comparing the mapping rate and on target rate between runs and samples, 2) verify and validate CNV calling using PiVAT, 3) identify limitations and improve on the specificity and sensitivity on CNV calls for MET, ERBB2, and MYC genes, and 4) benchmark PiVAT's CNV caller against other CNV callers: CNVkit and Control-FREEC with the same dataset.

EXPERIMENTAL FLOW

METHODS

Samples were sequenced using MGISEQ-2000 with a paired end, 2x100 read length protocol and adapter sequences were trimmed from the 3' ends of each read to create FASTQ files. BWA (Burrows-Wheeler Aligner) is implemented within the PiVAT pipeline to align FASTQ file sequences against the hg19 reference human genome and output BAM files. Sequencing run statistics are generated based on these alignment files which are then used to analyze variant and CNV calling. PiVAT's CNV calling method is based on double coverage normalization. This includes one per-sample normalization and one per-amplicon normalization with a minimum of 2 normal samples to be used as normalization references. For these analyses, we used 14 confirmed CNV negative samples as normal samples.

MET, MYC, and ERBB2 copy number was investigated to determine the sensitivity, specificity, precision, and accuracy of the overall run. These genes act as oncogenes and their activation by CNV amplifications are well-known biomarkers in cancer. A classification model was created to count the number of cases of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). With these classification counts, the sensitivity *(TP / [TP + FN]*, proportion of positive CNV calls compared to the total number of true CNV calls), specificity *(TN / [FP + TN]*, proportion of negative CNV calls compared to the total number without CNV calls), precision *(TP / [TP + FP]*, proportion of positive CNV calls compared to the predicted positive CNV calls), and accuracy (*[TP + TN] / [FP + FN + TP + TN]*, proportion of the classifier is correct) can be calculated.

BENCHMARKING CNV CALLING TOOLS

We wanted to compare the performance of PiVAT's CNV caller with 2 other CNV calling tools: CNVkit[2] and Control-FREEC[3]. These CNV callers were chosen because they are frequently used and publicly available. Paired end assembled PiVAT-filtered BAM files (PBAM) were used as input for each of these tools. PBAM files filter out poor quality reads from BAM files generated from BWA. This step serves to ensure that variant calling can be done with less background noise.

CNVkit uses on-target reads and nonspecific captured off-target reads to calculate $\log_2$ copy ratios across the genome for each sample. The off-target bins are taken from genomic positions between targeted regions. Both the on-target and the off-target locations are separately used to calculate the mean read depth and normalized to the control samples used as the reference. These are corrected for systematic biases to create a final $\log_2$ copy ratio table. CNVkit is primarily designed for use on hybrid capture sequencing data where off-target reads are prevalent, but it also has a Targeted Amplicon Sequencing protocol appropriate for our use case. We ran CNVkit with default parameters with a substitute blank file to serve as antitargets (off-target regions). This approach does not collect copy number information between targeted regions and does not attempt to normalize each amplicon at the gene level.

Control-FREEC uses a sliding window approach to calculate read count in non-overlapping windows with or without a control sample. If a control sample is available, the tool normalizes raw copy number profiles by using the control profile. Otherwise, it will normalize using GC content. To run Control-FREEC, we used the default settings with the inclusion of setting *mateOrientation = 0* for single end reads from PBAM files for the Control-FREEC Config file. Control-FREEC requires matching normal and tumor sample to run their CNV calling algorithm. Because our CNV positive samples do not have matching control samples, we chose to run the same CNV normal sample with each of the CNV positive samples.

The output generated from the tools was exported as report files with copy number ratios and copy number calls per sample. CNVkit and Control-FREEC make independent copy number calls at targeted amplicon regions specified by a supplied BED file. To account for this, we took the average copy number calls per gene and used that to compare ERBB2, MET, and MYC copy number levels by each method.

EXPLANATION OF DATASET

DNA INPUT STUDY

We wanted to understand the limit of detection for PiVAT at different DNA inputs by observing the overall statistics for coverage uniformity, mapping rate and the on-target rate. The dataset used for this study included a total of 60 CNV samples featuring 46 CNV positive and 14 CNV negative samples with inputs of 20ng, 40ng, and 60ng of DNA run using PiVAT.

CONCORDANCE WITH CNV CALLS

For this analysis, all 46 known CNV positive samples were tested using PiVAT's CNV caller to check the output copy number ratio is < 0.8 (copy number deletion) or > 1.2 (copy number amplification). PiVAT CNV output calls were compared against the expected copy number to calculate and plot the concordance and the Line of Best Fit.

Table 1 depicts DNA input amounts and copy number values for genes MET, MYC, and ERBB2 in the 46 samples used for CNV calling. CNV positive samples were further diluted to include a range of copy number for these analyses. We also used all CNV negative samples to serve as normal copy number 2 samples.

RESULTS

GENERAL STATISTICS FOR PIVAT

We analyzed the 46 CNV positive samples with 20ng, 40ng and 60ng DNA input to understand the limit of detection for PiVAT at different DNA inputs. Mean base coverage percentages, mapping rate, and on-target rates were nearly identical between the CNV positives samples at different DNA input amounts **(Figure 1A)**. All 46 samples grouped by DNA input displayed high coverage uniformity, with > 90% of sites having base coverage depth > 20% mean coverage for 20ng, 40ng, and 60ng of DNA input. A larger margin of error is seen with 60ng compared to 20ng and 40ng showing that lower DNA input is better for targeted sequencing but will not affect Pillar's sequencing ability to target a region of interest with a > 98% on-target rate. Input levels did not significantly affect the run statistics nor did they affect PiVAT CNV calling. A minimum of 90% is an industry standard acceptable performance level for each of these statistics. **Figure 1B** shows the performance of PiVAT with a mapping rate of 98.98±0.41% and an on-target rate of 98.39±0.45% for the 46 CNV positive cohort.

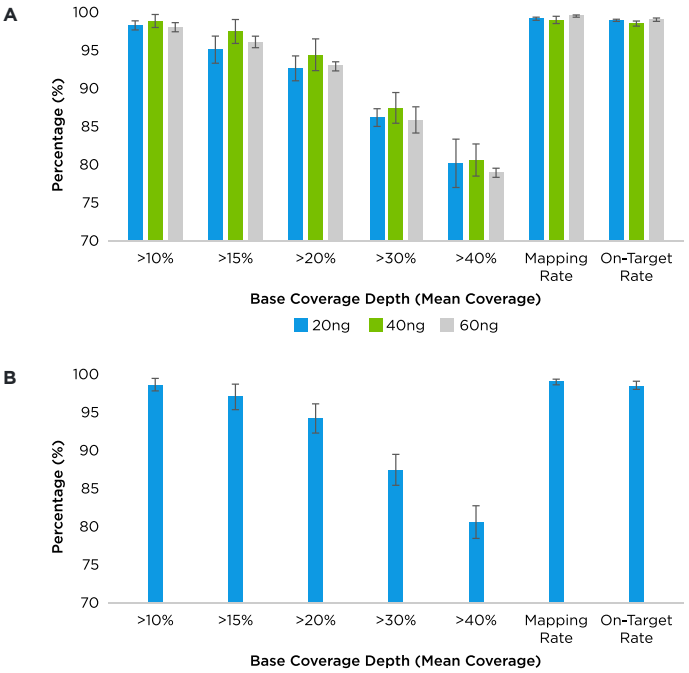

Figure 1 Overall statistics for coverage uniformity, mapping rate and on-target rate for PiVAT. A) The performance of PiVAT for the 46 CNV positive cohort. B) All CNV positive samples grouped by DNA input at 20ng, 40ng and 60ng.

PIVAT CNV CALLING PERFORMANCE

We focused on 3 genes—MET, MYC, and ERBB2—amplifications of which are known to be oncogenic biomarkers. Using samples containing CNV mutations involving these 3 genes, we validated the classification and measurement of PiVAT's CNV calls **(Table 2)**. PiVAT had a sensitivity of 97.83% and a specificity of 100%. Only 1 copy number amplification (CNA) was missed. This MET CNA had a copy number ratio of 1.18, below PiVAT's limit of detection (copy number 2.4n) when the expected copy number value is 3. No false positives were identified within CNV negative samples.

BENCHMARKING PIVAT AGAINST OTHER CNV CALLERS

We benchmarked PiVAT against two other CNV calling tools—CNVkit and Control-FREEC—and used the same 46 CNV positive cohort to compare performance and CNV calling capability. PiVAT and CNVkit used the same 14 CNV negative samples as a baseline for coverage normalization, while Control-FREEC required tumor-normal matched samples. Because our CNV positive samples do not have a matched normal, for each Control-FREEC comparison, we chose one sample at random to serve as the normal. Because of this Control-FREEC's performance may not have been ideal.

PiVAT more accurately estimated copy number when compared to CNVkit and Control-FREEC. PiVAT's R^2, measuring the degree of difference from expected copy number, was greater than CNVkit and Control-FREEC for MET, MYC and ERBB2, with values at 0.777, 0.6878, and 0.9649, respectively. In contrast, CNVkit and Control-FREEC were on average 77.38% and 164.04% worse, respectively **(Figure 2)**. For both CNVkit and Control-FREEC, many of the MET, MYC and ERBB2 CNV calls were seen at a much lower magnitude than the expected CNV call.

DISCUSSION

We have shown that Pillar's sequencing methods are reliable and accurate with the help of PiVAT's ability to detect the overall statistics of a run. Our coverage uniformity, mapping rate, and the on-target rate for the DNA input cohort and the CNV Positive sample group is shown to be > 98%, surpassing our acceptable 90% threshold. We have shown that lower DNA input is better for targeted sequencing and is more stable.

Our studies have also shown that PiVAT is a reliable tool for calling CNVs. PiVAT's sensitivity and accuracy was calculated to be > 97% and showed all CNV negative samples being called correctly as expected. PiVAT outperformed the other CNV calling tools CNVkit and Control-FREEC. The concordance plots showed PiVAT has a larger R^2 than the other CNV tools with a limitation to our study: we could not control the specific sample cohorts that both CNVkit and Control-FREEC require to accurately output copy number calls.

Sample ID	DNA Input Amount (ng)	MET Copy Number*	MYC Copy Number*	ERBB2 Copy Number*
Sample 1	40	4.5	9.5	
Sample 2	40	4.5	9.5	
Sample 3	20	3	5	
Sample 4	20	3	5	
Sample 5	40	3	5	
Sample 6	40	3	5	
Sample 7	60	3	5	
Sample 8	60	3	5	
Sample 9	20	3.47	6.41	
Sample 10	20	3.47	6.41	
Sample 11	40	3.47	6.42	
Sample 12	40	3.47	6.42	
Sample 13	60	3.47	6.43	
Sample 14	60	3.47	6.43	
Sample 15	20	3	5	
Sample 16	20	3	5	
Sample 17	20	3.47	6.41	
Sample 18	20	3.47	6.41	
Sample 19	40	3.47	6.42	
Sample 20	40	3.47	6.42	
Sample 21	40	3.5	7	
Sample 22	40	3.5	7	
Sample 23	40	3		5.33
Sample 24	40	3		5.33
Sample 25	40	3		5.33
Sample 26	40	2.6		4
Sample 27	40	2.6		4
Sample 28	40	2.6		4
Sample 29	40	2.34		3
Sample 30	40	2.34		3
Sample 31	40	2.34		3
Sample 32	40	2.186		2.5
Sample 33	40	2.186		2.5
Sample 34	40	2.186		2.5
Sample 35	40	3		5.33
Sample 36	40	3		5.33
Sample 37	40	3		5.33
Sample 38	40	2.6		4
Sample 39	40	2.6		4
Sample 40	40	2.6		4
Sample 41	40	2.34		3
Sample 42	40	2.34		3
Sample 43	40	2.34		3
Sample 44	40	2.186		2.5
Sample 45	40	2.34		3
Sample 46	40	2.34		3
CNV Negative	40	2	2	2

Table 1

*Portions of this table are left blank because the copy number value for MET, MYC, and ERBB2 genes are not known.

CNV Gene	FN	FP	TN	TP	Sensitivity	Specificity	Precision	Accuracy
MET	1	0	14	45	97.83%	100.00%	100.00%	98.33%
MYC	0	0	14	20	100.00%	100.00%	100.00%	100.00%
ERBB2	0	0	14	26	100.00%	100.00%	100.00%	100.00%

Table 2 Classification and measurement of PiVAT CNV calls for ERBB2, MET, and MYC. CNVs in 60 samples (46 CNV positive and 14 CNV negative) were analyzed by identifying the classification counts of FN: False Negative, FP: False Positive, TN: True Negative, and TP: True Positive to calculate the sensitivity, specificity, precision, and accuracy of PiVAT's CNV caller. Sensitivity represents the proportion of positive CNV calls compared to the total number of true CNV calls, specificity is the proportion of negative CNV calls compared to the total number without CNVs, precision is the proportion of positive CNV calls compared to the predicted positive CNV calls and accuracy is the proportion of the classifier is correct

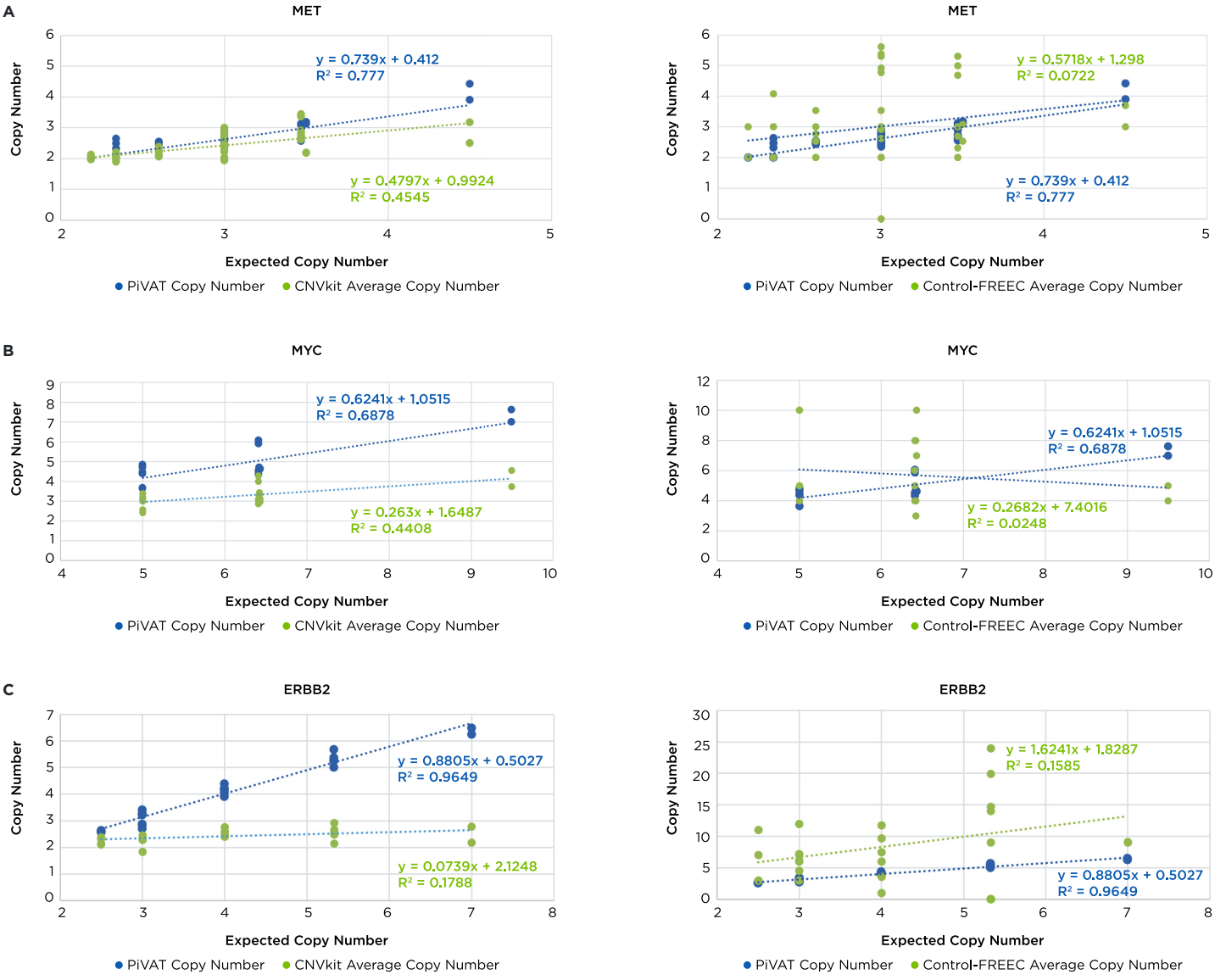

Figure 2 Concordance between expected and observed copy numbers of MET, MYC, and ERBB2 using PiVAT, CNVkit, and Control-FREEC. Copy number comparisons between A) MET, B) MYC and C) ERBB2 genes along with Line of Best Fit and the coefficient of determination (R^2) calculated for PiVAT (Blue), CNVkit (Green) and Control-FREEC (Green). Values at > 0.80 for both the y-intercept from the Line of Best Fit and R^2 are acceptable for the performance of each of the CNV callers.

REFERENCES

1. Soave A, Kluwe L, Yu H, *et al*. Copy number variations in primary tumor, serum and lymph node metastasis of bladder cancer patients treated with radical cystectomy. *Sci Rep*. 2020;10:21562. doi:10.1038/s41598-020-75869-x
2. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput Biol*. 2016;12(4):e1004873. doi:10.1371/JOURNAL.PCBI.1004873
3. Boeva V, Popova T, Bleakley K, *et al*. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423-425. doi:10.1093/BIOINFORMATICS/BTR670

Learn more at
pillar-biosciences.com

PILLAR BIOSCIENCES

For Research Use Only.

Pillar Biosciences, Inc.
9 Strathmore Rd
Natick, MA 01760
(800) 514-9307
techsupport@pillar-biosciences.com