Assessing viral variant detection algorithms to improve characterization of gene therapy products

Hiral Desai, Yanfei Zhou, Ph.D., Bradley Hasson, Alexandra Bridgeland, William Dolan, Amber Overgard
MilliporeSigma, Rockville, USA

Introduction

Identity testing is an evolving requirement of regulatory agencies to characterize biologics intended for viral and gene therapy products (Figure 1). Next generation sequencing (NGS) can be used to characterize and confirm the identity of viral vectors delivering genetic material to affected cells in a patient by creating a full genetic profile of all nucleic acids contained within the test sample.

It is important to characterize viral vectors to ensure they do not contain variants which can negatively impact patient outcomes. Downstream bioinformatics analysis must capture true positive variants and limit spurious results to establish sequence identity and purity of the expression vectors while maintaining sensitivity and specificity. In addition, all steps of a bioinformatics workflow must be carefully examined to configure necessary requirements for optimal results. Advancements in best practices and standards for viral variant detection is critical to ensure the safety of patients and meet the expectations of regulatory guidance.

The analysis of over 10 variant callers and other bioinformatics tools for viral variant detection is discussed to better understand how the outcomes can be applied to improve characterization of gene therapy products.

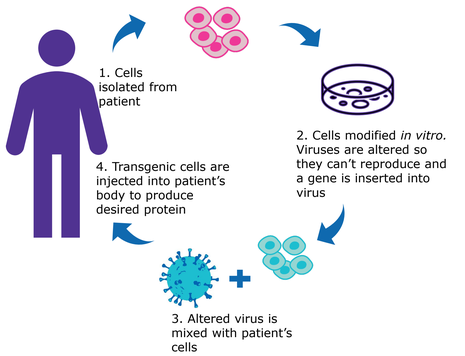

Figure 1: Example of gene therapy process. In step 1, cells are isolated from patient. The cells are then modified *in vitro* in step 2. In step 3, an altered virus is mixed with the cells before being injected back into the patient in step 4. NGS algorithms can identify purity of viral vector used for gene therapy.

Materials and Methods

Synthetic Data Generation

Synthetic data using 20 different variants at known positions within five categories of mutations were created using a partial viral genome (Table 1). The categories include deletions (del), indels, insertions (ins), multi-nucleotide variants (mnvs), and single nucleotide polymorphisms (snps) Each variant category with known variants were generated at 100%, 50%, 20%, 10% and 1% variant frequencies to examine sensitivity and specificity. Three sequencing replicates were generated per variant category and frequency on the MiSeq™ system and NS2000 instrument for a total of 156 datasets to test bioinformatics variant calling algorithms.

REFERENCE NAME	VARIANT TYPE	POSITION	CHANGE
del	deletion	197	100 bp
del	deletion	1261	240 bp
indel	insertion	800	+4C
indel	snp	1327	G>T
indel	snp	1329	A>T
indel	snp	1335	T>A
indel	deletion	1333	1 bp
indel	deletion	1998	2 bp
ins	insertion	14	12 bp
ins	insertion	1058	28 bp
mnv	insertion	880	4 bp
mnv	ins	1326	28 bp
mnv	del	1888	40 bp
mnv	del	1998	3 bp
snp	del	1	1 bp
snp	snp	50	A > T
snp	snp	920	T > G
snp	snp	1531	T > C
snp	snp	1944	C > A
snp	snp	1973	C >T

Table 1: List of references with known variants and position information which were used for creating synthetic data for bioinformatics analysis. Column 1 indicates the reference name. Column 2 indicates variant type such as snp, deletion, insertion, column 4 indicates position information in reference, and column 5 shows the variant.

Bioinformatics Analysis

Over 10 open-source variant callers were utilized to examine viral variant detection for 156 datasets. Figure 2 shows an example bioinformatics workflow for identity analysis. Each step must be assessed to identify optimum parameters for viral variant detection and the variant caller used must identify all known variants without introducing false positive results.

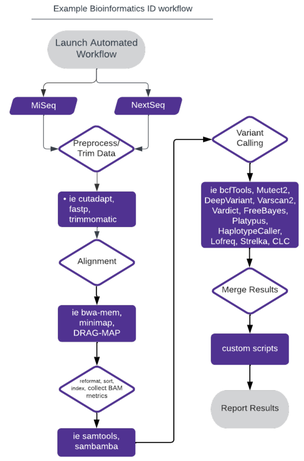

Figure 2: Example bioinformatics identity detection workflow which starts with sequencing data. The reads are preprocessed, aligned to a reference and the output is used for variant calling with different open-source tools.

Different statistical measures were calculated for all datasets (Table 2) and used to create graphs representing results.

Statistic	Definition	Equation
TP	Number of true positive variants	–
FP	Number of false positive variants	–
FN	Number of false negative variants (missed variants)	–
PPV/Precision	positive predictive value or fraction of correctly identified variants. Closer to 1 = better.	$\frac{TP}{(TP+FP)}$
FDR	False discovery rate. Closer to 0 = better	$1 - PPV$
Sensitivity/Recall	Completeness or measure of how well variant caller could detect variants at all frequencies. Closer to 1 = better.	$\frac{TP}{(TP+FN)}$
F1 Score	Accuracy; harmonic mean of PPV and Recall. Allows comparison based on single metric. Closer to 1 = better.	$\frac{2 * PPV * Recall}{(PPV + Recall)}$

Table 2: Equations used to analyze results. Column 1 represents the statistical measure: true positives, false positives, false negatives, precision, false discovery rate, sensitivity, and F1 scores. The F1 scores allow comparisons between results using one metric. Ideally the value should be close to 1.

Results and Discussion

Results from analyzing snps are displayed as an example of analysis conducted on all variant categories. The same analysis was conducted on indels, mnvs, insertions, and deletions. Table 3 shows that Platypus, Mutect2, Varscan, and Vardict had similar average F1 scores for snps across all expected variant frequencies, however discerning results at each frequency level can determine most suitable variant caller for low level variant detection.

Variant Caller	Variant Type	Ave TP	Ave FP	Ave FN	Ave PPV	Ave FDR	Ave Sensitivity	Ave F1 Score
BCFTools	snp	2.4	0	0.48	0.48	0	0.4	0.44
Haplotype Caller	snp	3	0	0.6	0.6	0	0.5	0.55
Platypus	snp	4	0	0.8	0.8	0	0.67	0.73
Mutect2	snp	4	0	0.8	0.8	0	0.67	0.73
Varscan	snp	4	0	0.8	0.8	0	0.67	0.73
Strelka	snp	2	0	0.4	0.4	0	0.33	0.36
Vardict	snp	4	0	0.8	0.8	0	0.67	0.73
FreeBayes	snp	3	0	0.6	0.6	0	0.5	0.55
DeepVariant	snp	0	0	0	0	0	0	0
LoFreq	snp	2	0	1.6	0.6	0	0.33	0.40

Table 3: Average results of TP, FP, FN, PPV, FDR, Sensitivity and F1 score of all variant callers across all variant frequencies for snps are shown. Platypus, Mutect2, Varscan, and Vardict had the highest F1 score of 0,73 for snp detection.

Figure 3 shows individual F1 scores for snps across all expected variant frequencies. The same analysis was conducted on indels, mnvs, insertions, and deletions. Only lofreq, mutect2, platypus, vardict and varscan are able to resolve variants at the 1 percent frequency in addition to all other variant frequencies while maintaining sensitivity and specificity. These variant callers can be used to continue further bioinformatics analysis to determine optimal parameters for viral variant detection.

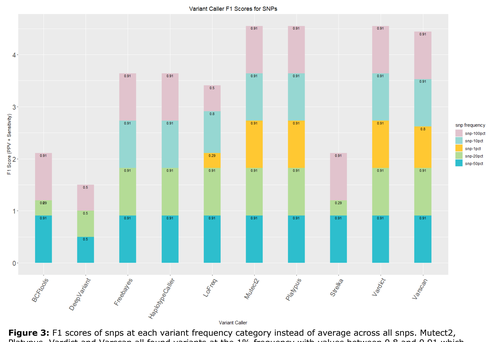

Figure 3: F1 scores of snps at each variant frequency category instead of average across all snps. Mutect2, Platypus, Vardict and Varscan all found variants at the 1% frequency with values between 0.8 and 0.91 which means sensitivity and specificity are maintained.

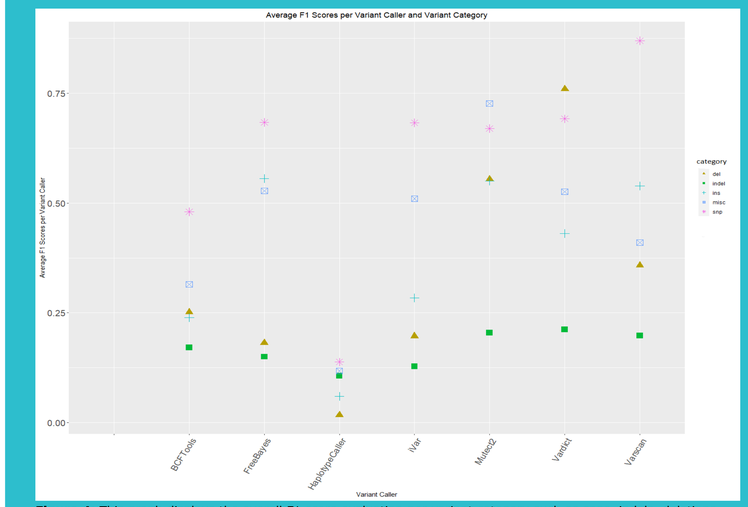

Figure 4: This graph displays the overall F1 score evaluation per variant category such as snps, indels, deletions, insertions, and mnvs. Results show that each variant caller is optimized to detecting specific types of variants. Varscan had the highest F1 Score for snps while vardict is optimized for discovering long insertions and deletions (>20 bp). All variant callers had difficulty with the designed indel dataset at each variant frequency and F1 scores were 0.25 or less. Due to design flaw; this experiment will be repeated with data better representing indels to determine optimal variant caller for indels.

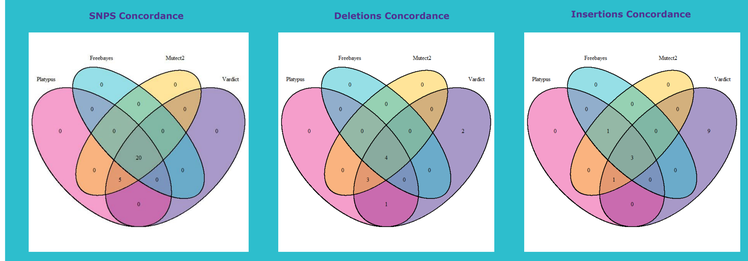

Figure 5: These diagrams show concordance among playtpus, freebayes, mutect2, and vardict for discovering SNPs, deletions, and insertions. All variant callers can find most expected snps at all frequencies without discovering false positive results, indicted by the overlapping value of 20. Results are variable for deletions and insertions of which vardict finds most expected deletions and insertions across all expected variant frequencies. There are only 4 overlapping and 3 overlapping insertions found deletions found among all variant callers. Because of the varying level of concordance depending on variant type, a combination of variant callers must be used to successfully detect multiple variant types.

Conclusion

- NGS can be used to detect viral variants at the 1% frequency depending on variant caller used while maintaining sensitivity and precision.
- Vardict is the best suited for discovering long insertions and deletions as it found all expected variants at all variant frequencies while, all other variant callers found 0-2 expected variants at each variant frequency.
- Vardict, Mutect2, Platypus, Varscan can discover snps at all variant frequencies and have F1 scores close to 1. They can be used for continuing analysis or a main variant caller for identity testing.
- Indel datasets were inconclusive due to the placement of the indels within the datasets. Multiple snps located 1 bp from other snps were used for analysis and variant callers could not discern the 1 bp difference. Indels were also placed towards beginning and end of reference which is already known to be problematic for many bioinformatics tools. This analysis will be repeated.
- Multiple variant callers optimized for discovering different variant types must concordantly be used to improve variant calling.

SigmaAldrich.com/BiosafetyTesting

BioReliance®
Contract Testing Services