DNASTAR
SOFTWARE FOR LIFE SCIENTISTS

A Next-Gen Sequencing Software Workflow for Gene Panel Validation Control

Matthew Keyser[1], Kerri Phillips[1], Timothy Durfee PhD[1,2], Jacqueline Carville[1], Thomas Schwei[1], Amber Pollack-Berti PhD[1], Daniel Nash[1], Jennifer Stieren[1], Schuyler Baldwin[1], Richard Nelson PhD[1], Kenneth Dullea[1], John Schroeder[1], Pavel Pinkas PhD[1], Guy Plunkett III PhD[1,2], Frederick Blattner PhD[1,2,3]

Affiliations:
[1] DNASTAR, Inc., Madison, Wisconsin, USA
[2] University of Wisconsin, Department of Genetics, Madison, Wisconsin, USA
[3] Scarab Genomics LLC, Madison, Wisconsin, USA

Abstract

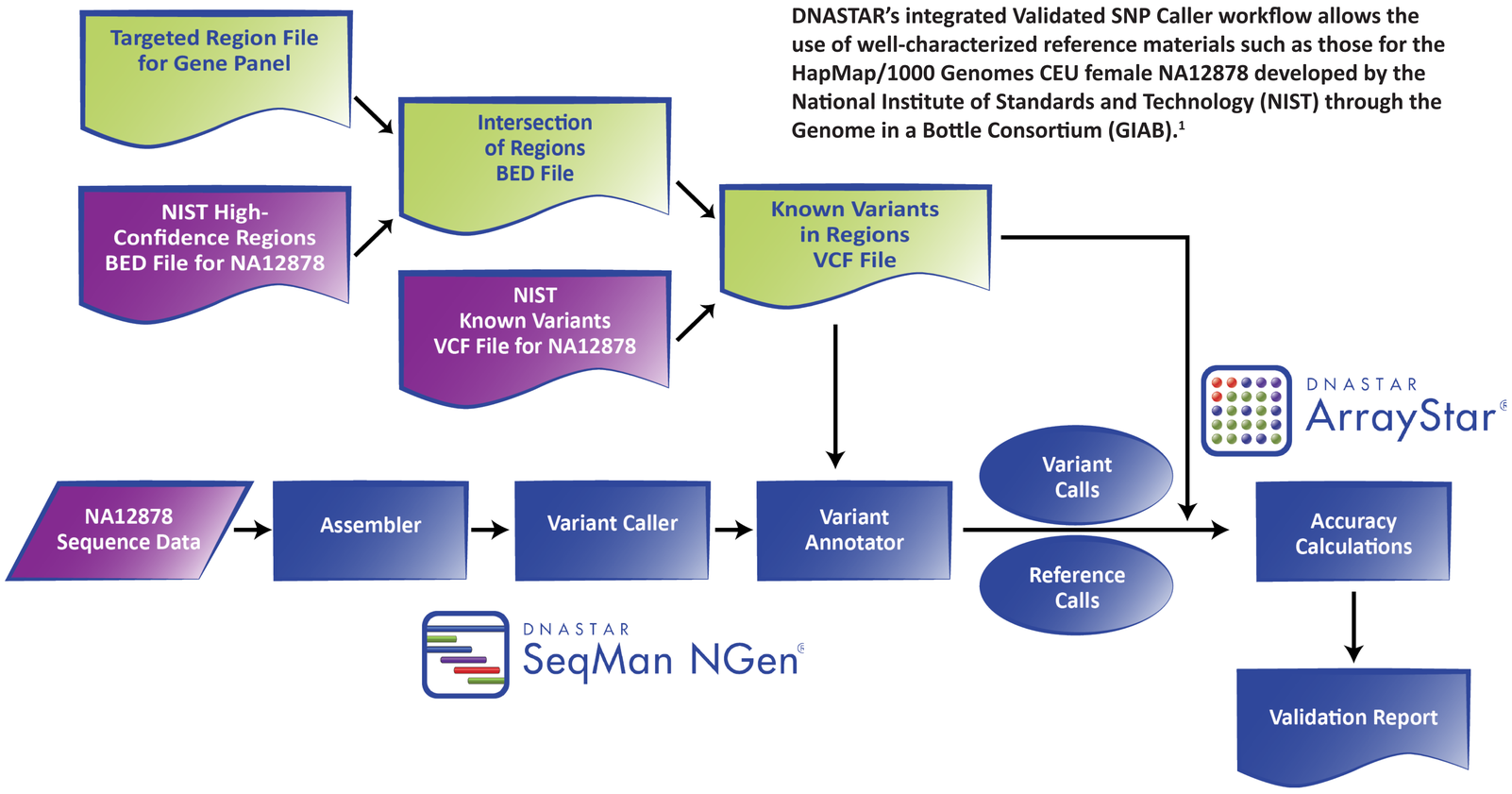
DNASTAR offers an integrated suite of software for assembling and analyzing sequence data from all major next-generation sequencing platforms. DNASTAR software supports a variety of key workflows on a desktop computer. Optional, on-demand computing power is also available through DNASTAR Cloud™ applications. A new Gene Panel with Control Validation workflow supports several types of data sets, including Ion AmpliSeq™ Comprehensive Cancer Panel, Illumina® TruSight™ Cancer Panel, and custom gene panels. This workflow facilitates the evaluation of, not just the accuracy of the software's variant calling, but also the efficacy of the gene panel targeting.

The Gene Panel with Control Validation workflow allows a well-characterized control to be processed and analyzed in parallel with test samples. In the context of the known variants that are expected to be called, the accuracy of the variant calling results from the control sample is then calculated, providing a measure of each run's performance. Recommended controls include the reference materials provided by the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIAB) consortium.

Gene panel targeting accuracy is determined by multiple factors, including the specificity of primers and probes used for gene panel design, efficiency of the sequencing technology, accuracy of the assembly, accuracy of the variant calling, and the filters applied. DNASTAR's SeqMan NGen® and ArrayStar® programs provide an accurate alignment algorithm and variant caller. They then utilize a validated SNP set for the control sample in the form of a VCF file along with a BED or Manifest file, which specifies targeted regions, to calculate variant-calling sensitivity, specificity, and accuracy. By utilizing this workflow, users can ultimately validate their entire process to verify that their targeted variants are being identified.

Workflow

DNASTAR's integrated Validated SNP Caller workflow allows the use of well-characterized reference materials such as those for the HapMap/1000 Genomes CEU female NA12878 developed by the National Institute of Standards and Technology (NIST) through the Genome in a Bottle Consortium (GIAB).[1]



Calculations

Specific metrics are used to calculate Validation Report results.

Each position in the targeted region is classified into one of four categories:

1. True Positives (TP). Called variants with a corresponding position in the VCF file.
2. False Positives (FP). Called variants without a orresponding position in the VCF file.
3. True Negatives (TN). Called reference bases without a corresponding position in the VCF file.
4. False Negatives (FN). Called reference bases with a corresponding position in the VCF file.

The counts of each class are then used to calculate the False Positive Ratio, False Discovery Rate, Sensitivity, Specificity and Balanced Accuracy (Table 1). Sensitivity measures the proportion of true positives that are correctly identified. Specificity measures the proportion of true negatives that are correctly identified.

False Positive Ratio	**FPR=FP/(FP+TN)**
False Discovery Rate	**FDR=FP/(FP + TP)**
Sensitivity	**Sens. = TP/(TP+FN)**
Specificity	**Spec. = TN/(FP +TN)**
Balanced Accuracy	**(Sens. + Spec.)/2**

Table 1. Calculations used.

Compare Results

Use the Validation Report to compare results obtained from using different technologies or processes.

Data	Sensitivity	Specificity	Balanced Accuracy
Illumina exome[2]	99.655%	99.998%	99.826%
Ion Torrent exome[3]	97.855%	99.723%	98.789%

The table above shows the results of the Lasergene Validation Control Workflow when conducted using data sets from two different NGS technologies. Calculations were done on variants detected with a minimum depth of coverage of 20 and PNotRef > .90. PNotRef is the probability that the called base is not the homozygous reference base. This value is used as a minimum threshold for counting positives in the validation control workflow. For this analysis, whole exomes were analyzed from both an Illumina paired-end data set[2] and an Ion Torrent data set.[3]

Improve Process

The Validation Report illustrates if targeting efficiency and variant detection sensitivity are at desired levels. If not, Lasergene SeqMan Pro can help identify where to focus improvements.

Coverage of Targeted Regions Table. A table of genomic coordinates specified in a BED file that provides coverage information for all gene targets.

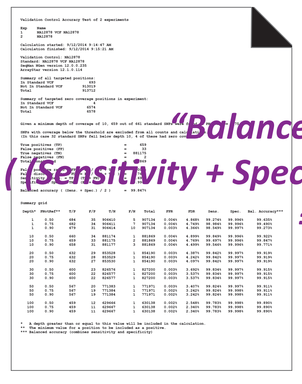
Enrichment Report. Provides a series of summary statistics pertaining to the quality of the region capture and assembly processes. This report provides such information as the total length of the targeted reference sequence and the mean region coverage depth.

Missing SNP Report. Identifies variants that are known to be in the control reference material at specific locations but were not detected during a run.

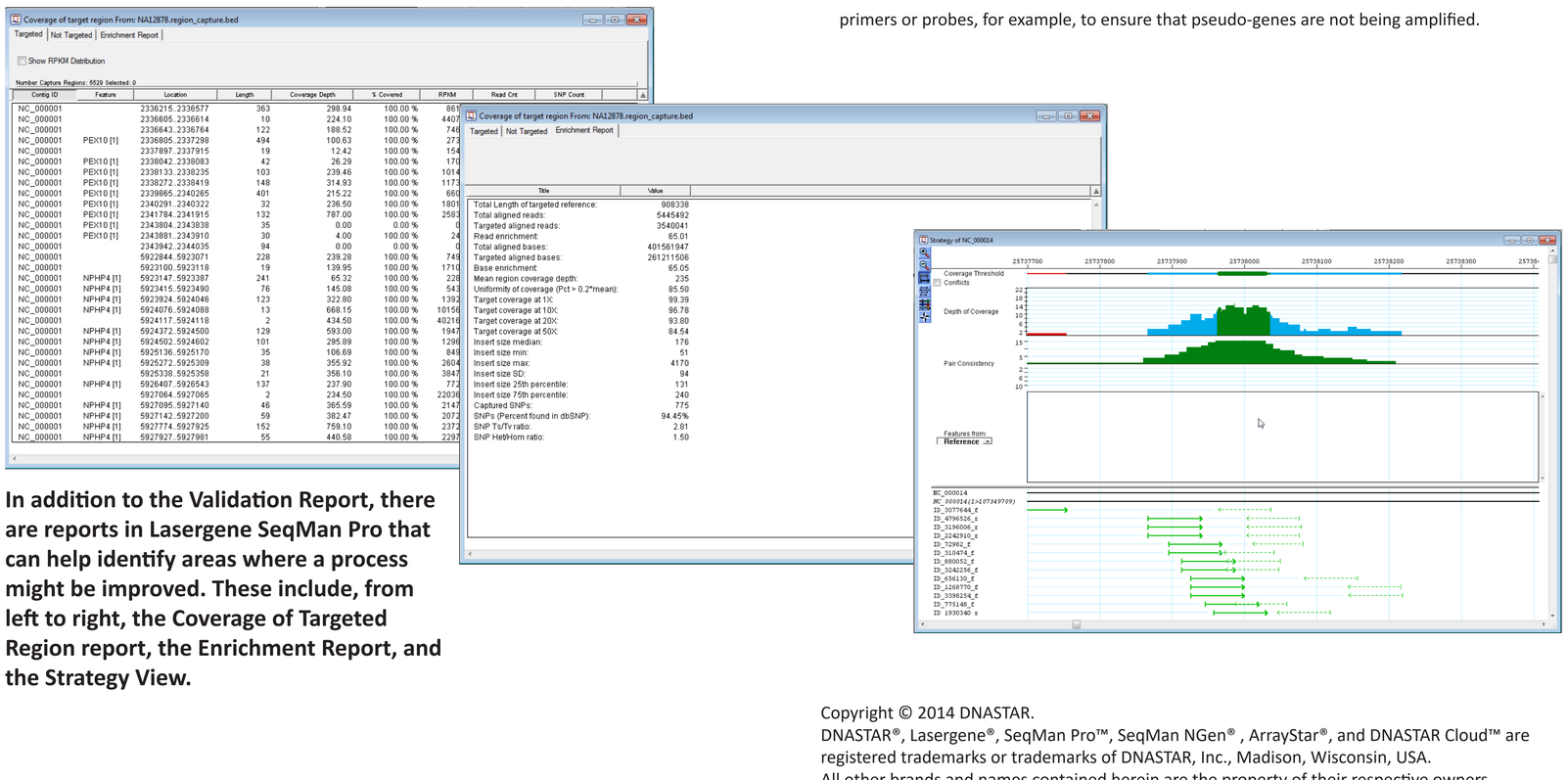
BED or Manifest file filtering. By default, SNPs and small indels that are aligned outside of the targeted region are filtered out of the SNP report. To get an assessment of how specific the targeting is, all variants can be displayed. This can help with analyzing the effectiveness of the targeting and support decision making regarding possible adjustments needed to the primers or probes, for example, to ensure that pseudo-genes are not being amplified.

Validation Report

Automatically generated statistical report details the assembly sensitivity, specificity, and accuracy.



"Balanced accuracy ((Sensitivity + Specificity) / 2) = 99.835%"



In addition to the Validation Report, there are reports in Lasergene SeqMan Pro that can help identify areas where a process might be improved. These include, from left to right, the Coverage of Targeted Region report, the Enrichment Report, and the Strategy View.

References

1. Zook, J. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology 32, 246-251. (2014)

2. Illumina paired end data set produced by the Garvan Institute was obtained from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/technical/garvan_data/,

3. Ion Torrent data set was obtained from the Ion Community website, http://ioncommunity.lifetechnologies.com.

For Research Use Only.
Not for use in diagnostic procedures.