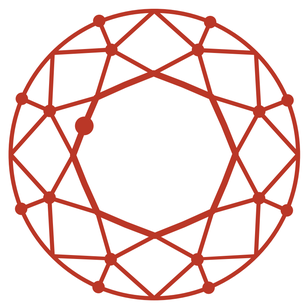
Strategies for calculating variant confidence by combining sequencing results



CLARITAS GENOMICS

Niru Chennagiri, Daniel Lieber, Timothy Yu, John Thompson

Abstract

Clinical labs using High Throughput Technology (HTS) are required to confirm all variants they report using a companion technology – usually Sanger sequencing. As whole exome sequencing becomes increasingly popular, labs are faced with having to confirm a large number of variants which increases cost and turn around time. At Claritas Genomics, we sequence a sample using two HTSs – Illumina NextSeq and Ion Torrent Proton. We have developed methods to combine the two calls into a consensus call and assign confidence levels. We assign four levels of confidence based on Positive Predictive Value (PPV) or precision of the consensus calls calculated using NA12878 NIST reference dataset. About 85% of the total calls are concordant between the two technologies and are assigned the highest level of confidence. The calculated PPV for such variants is 100%. These variants can be reported without additional Sanger Sequencing. The rest of the variants are assigned confidence levels based on their PPV and are either dropped or Sanger confirmed. The confidence levels give an easy way of prioritizing variants to confirm and significantly reduce turn around time.

Methods

Combining Results from Orthogonal Sequencing

The samples were sequenced using two different sequencing technologies - Illumina NextSeq and Ion Torrent Proton. Variants were called on the NextSeq data using GATK Best Practices pipeline and on the Proton using Torrent Suite 4.4 software. The resulting Variant Call Format (VCF) files were processed using Combinator software developed by Claritas. The Combinator makes a consensus call for each variant detected by either platform and assigns a category (Table 1) to the variant based on the call status of the variant in each platform.

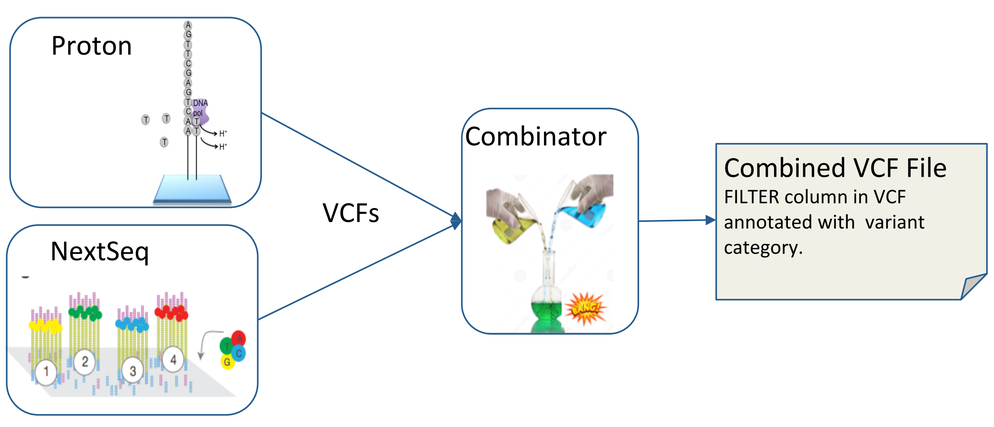


Figure 1: Combinator software combines VCFs from Proton and NextSeq and creates a combined VCF with consensus calls and an assigned variant category.

Assigning Confidence Levels.

Combinator assigns a confidence level to each variant category calibrated using the Genome In A Bottle (GIAB) NA12878 reference dataset. To calculate the confidence level, we sequence the NA12878 using the two platforms and combine the variant calls as described above. We compare the combined to the reference VCF from GIAB using the software Comparator developed by Claritas. For each category, we then calculate the Positive Predictive Value (PPV) defined as the fraction of True Positive calls out of all the Positive calls. The confidence level is assigned as one of Orthogonally Confirmed, Likely True Positive or Likely False Positive based on the PPV for that category (Table 1)

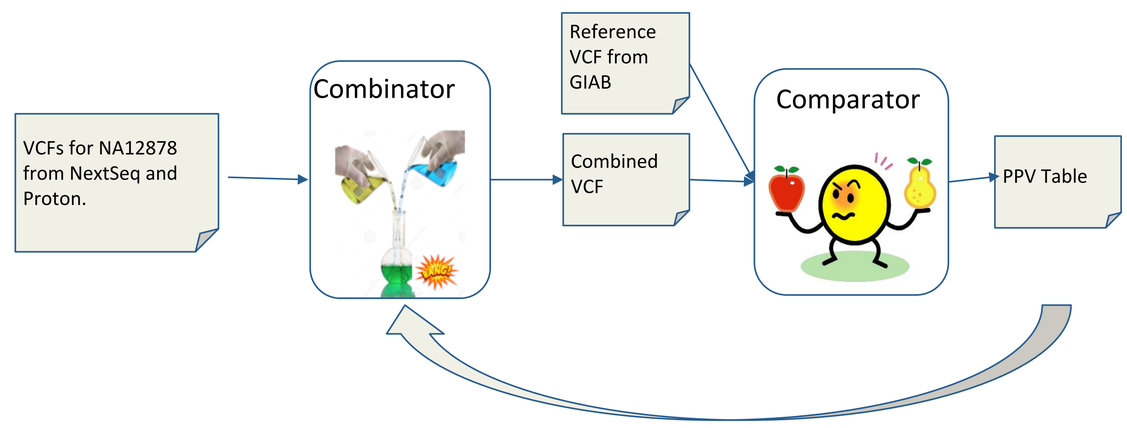


Figure 2: Confidence levels are assigned to each variant category based on the PPV calculated using reference GIAB dataset.

Results:

Table 1: PPV table and Confidence Levels for each variant category. 92% of the variants are confirmed orthogonally on two platforms.

Variant Category	PPV	Confidence Level	Fraction Of Total
proton-illumina-match-indel	100%	Orthogonally Confirmed	92%
proton-illumina-match-snp	100%		
proton-illumina-not-PASS-match-indel	100%		
proton-illumina-not-PASS-match-snp	100%		
illumina-snp-proton-indel	NA	Likely True Positive	7%
no-match-indel-call	NA		
illumina-snp-proton-nocov	99%		
no-match-snp-call	98%		
illumina-indel-proton-ref	95%		
proton-snp-illumina_nocov	94%		
illumina-snp-proton-ref	93%		
illumina-indel-proton-nocov	92%		
illumina-not-PASS-snp-proton-nocov	46%		
proton-indel-illumina_nocov	NA		
illumina-not-PASS-indel-proton-nocov	NA		
illumina-snp-not-PASS-proton-indel	NA		
proton-snp-illumina-indel-not-PASS	NA		
illumina-not-PASS-indel-proton-ref	22%	Likely False Positive	1%
illumina-not-PASS-snp-proton-ref	38%		
proton-indel-illumina-ref	7%		

Table 2: Sensitivity and specificity of the combined VCF data by comparing with NIST reference in the region defined by RefSeq coding exons + 10 bp intersected with NIST called regions.

Variant Type	SENSITIVITY	SPECIFICITY (FP/MB)	PPV
All Variants			
SNV	99.4%	4.6	99.3%
Indel	92.0%	1.4	92.2%
Excluding Likely False Positive			
SNV	99.2%	1.5	99.8%
Indel	92.0%	0.7	96.1%

Discussion:

Table 1 shows that greater than 90% of the variants are detected on both platforms and can be reported without Sanger confirmation. Of the others, ~1% of the variants fall in the category of Likely False Positives with very low PPVs and can be disregarded from further consideration. Table 2 shows the sensitivity of the combined VCF over the whole exome region. We see over 99% sensitivity for SNVs and 92% for INDELs when we consider all variants. If we remove the category Likely False Positive, we see >50% reduction in False Positives with minimal change in sensitivity.

Orthogonal sequencing as described here helps with reducing the Sanger burden and also helps with prioritizing the variants both of which result in reduced Turn Around Time (TAT).

References:

1. Zook, J.M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251 (2014).
2. GATK Best practices paper

See other Claritas Genomics posters:
1619, 1981, 2070, 2071, 2085

CLARITAS GENOMICS INC.
99 Erie Street | Cambridge, MA 02139

clientservices@claritasgenomics.com
p: 617.553.5880 p: 855.373.9003 f: 617.553.5842