Benchmarking of Strand NGS variant caller using a whole genome sample NA12878 and data from genome in a bottle consortium

Ramesh Hariharan, Rohit Gupta*, Pallavi Gupta, Aishwarya Narayanan, Somak Aditya, Shanmukh Katragadda, Vamsi Veeramachaneni

Strand Life Sciences, Bangalore India. *Contact: rohit@strandls.com

Abstract

Many algorithms are developed for variant calling, however they differ on multiple variant call predictions. Here we present Strand NGS variant calling approach and benchmarking results on a whole-genome sample NA12878, comparing variant calls with those from GATK UnifiedGenotyper. Strand NGS and GATK identified a total of 6,393,054 and 6,105,466 variants respectively with very similar Het/Hom and Ti/Tv ratios. We observed a high overlap (98%) in the variant calls after filtering based on quality metrics, making Strand NGS and GATK very similar. We also observed a sensitivity of 84.28 and specificity of 99.9966 on a GCAT data using confident call set from genome in a bottle consortium.

Problem Statement and Challenges

- Variant calling algorithms compare the nucleotides present on aligned reads against the reference, at each position. Based on the distribution of As, Ts, Gs, and Cs at that position, and the likelihood of error, a judgement is made for the existence of a variant. Some issues that must be handled by variant detection algorithms are mentioned below:
  - Quality of base calls
  - Mapping quality of reads
  - Depth of coverage
  - Homopolymer
  - Ploidy

Strand NGS Variant Calling Approach

- Pre-processing step - Likely variant locations are determined based on the following requirements:
  - Reads coverage must exceed a user-defined threshold
  - Variant coverage must also exceed a user-defined threshold
- At every location that is declared significant in the previous step, the Bayesian variant calling algorithm is applied to identify the most likely genotype, and to report the variant if detected
  - First prior probability of each genotype at every location is calculated taking into account parameters such as reference base, heterozygosity rate, hom to het ratio, InDel to substitution ratio, T_i to T_v ratio
  - The algorithm then selects the genotype that maximizes the posterior probability, defined by the probability of a genotype G given the observed data D. This is easily computed under Bayes principle as

$$P(G/D) = \frac{P(D/G).P(G)}{P(D)} \quad (1)$$

Benchmarking Approach

- Strand NGS and GATK variant callers were run on NA12878 whole-genome data using their default parameters
- GATK processed (after local realignment and base quality score recalibration) BAM file is used as an input
- While GATK uses a variant quality score recalibration (VQSR) step to filter likely false positive SNPs, Strand NGS uses intuitive filters on variables like supporting read %, coverage, score, strand bias and other PV4 biases

Results: Number of Detected Variant

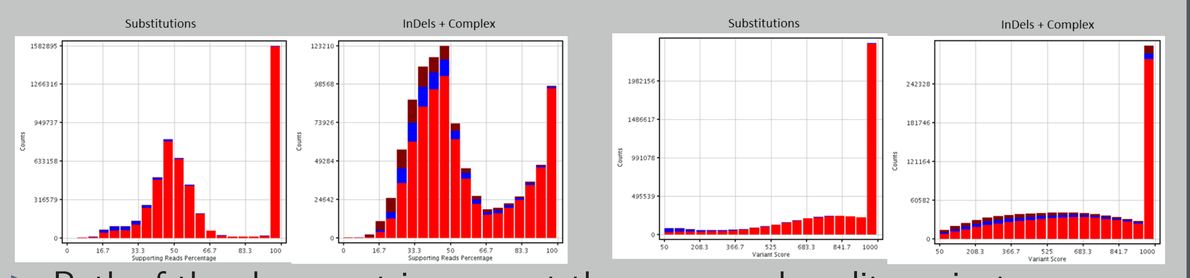
- Total number of variants detected by Strand NGS and GATK on NA12878:

	Strand NGS	GATK
Total variants	6,393,054	6,105,466
Substitutions	5,199,005	5,085,477
Insertions	467,267	468,203
Deletions	549,043	503,689
Complex variants	177,739	48,097
Het/Hom	2.55	2.08
T_i/T_v	1.92	1.93

- Of the 6,393,054 variants identified by Strand NGS, 5,948,107 variants (5,005,771 substitutions and 942,336 InDels and complex) are identified by GATK also, producing an overlap of 93%.
- There are a total of 444,947 variants (193,234 substitutions, 251,713 InDels and complex) uniquely identified by Strand NGS and a total of 157,359 variants (79,706 substitutions, 77,653 InDels and complex) uniquely identified by GATK.

Results: Characteristics of Overlapping Variants

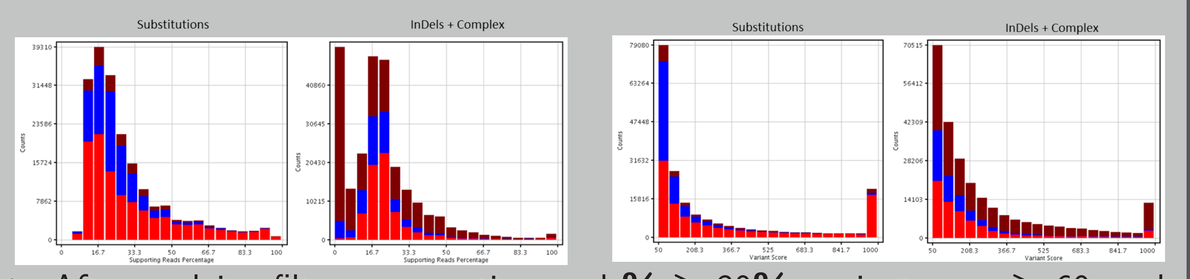
- Supporting read % and score is shown below



- Both of the above metrics suggest these are good quality variants

Results: Variants Uniquely Identified by Strand NGS

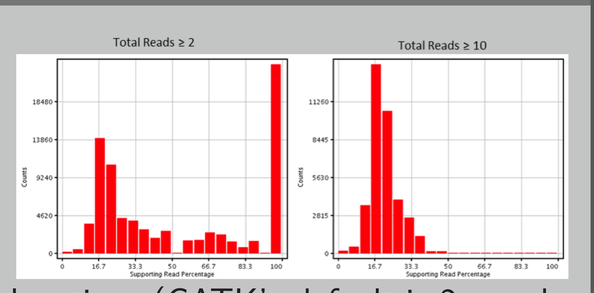
- Supporting read % and score for 444,947 variants is shown



- After applying filters: supporting read % $\geq$ 30%, variant score $\geq$ 60, and strand bias $\leq$ 50, only 145,324 variants were left in this category, making the overlap between Strand NGS and GATK as **98%**

Results: Variants Uniquely Identified by GATK

- Supporting read % for 157,359 variants is shown
- 21% of these are novel according to dbSNP138, indicating these may be likely false positives



- Further they may be at low coverage locations (GATK's default is 2, much lower than Strand NGS default of 10), low supporting read %, high strand bias, low variant score, presence of other PV4 biases etc. For instance supporting reads % shows that many of these variants can be filtered out due to relatively bad quality

Results: Benchmarking using genome in a bottle data

- Strand NGS variant caller is also benchmarked against genome in a bottle data
- Variant calling was done on GCAT Illumina 100bp 30x paired-end data
- BWA-mem based alignment output was used as input for Strand NGS and GATK variant caller

Algorithm	GIB Sensitivity	GIB Specificity	T_i/T_v
Strand NGS	84.28	99.9966	2.004
GATK UG	85.21	99.9975	2.148

Conclusion

- We demonstrated a high overlap (98%) in the filtered variant calls by Strand NGS and GATK, making them very similar for most practical purposes. Strand NGS provides an intuitive way to filter potentially false positive variants using quality metrics like dbSNP presence, supporting reads %, variant score, total read coverage, strand bias and other PV4 biases. Further Strand NGS also compares well with GATK when evaluated on GCAT Illumina 30x data using confident variant call set from genome in a bottle (GIAB) consortium