strandngs

Streamlining NGS Data Management & Analysis

Strand NGS Variant Caller

A Benchmarking Study

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

Abstract

Background

To realize the potential and promise of Next Generation Sequencing (NGS) technology towards research and clinical applications, computational approaches are essential to call variants and translate them into actionable knowledge. Many algorithms are developed for variant calling, however they differ on multiple variant call predictions. In this paper, we'll present Strand NGS (formerly AvadisNGS) variant calling approach and benchmarking results on a whole-genome sample NA12878 and compare our variant calls with those from GATK UnifiedGenotyper.

Results

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 (93%) in these variant calls with overlap in fact increasing to ~98% after filtering based on quality metrics, making Strand NGS and GATK very similar. Assessment of different quality metrics like supporting reads %, variant score, coverage, strand bias and other PV4 biases, for overlapping versus uniquely identified variants, provides a useful way to filter the likely false positives.

Conclusions

Due to several issues with the raw sequencing data, variant calling is still a challenging problem. Although numerous open-source and proprietary algorithms are available, assessing how well these algorithms perform on different data sets is not trivial. The benchmarking results presented in this paper suggest that Strand NGS variant caller is a powerful and flexible approach to call variants and provides a visually appealing way to assess their quality using a variety of metrics. The results compare very well with GATK, which is one of the widely used variant callers.

1. Introduction

Most organisms within a particular species differ very little in their genomic structure. A single nucleotide variation between two complementary DNA sequences corresponding to a specific locus is referred to as a SNP or Single Nucleotide Polymorphism and is the most common variation observed within a species. This variation could be due to insertion, deletion or substitution of a nucleotide. The variations are referred to as allele changes at the specific chromosomal loci.

Typically, SNPs commonly observed in a population exhibit two alleles – a major allele, which is more prevalent, and a relatively rarely occurring minor allele. Such variations, if they occur in genic regions, can result in changed phenotypes and may even cause disease. Sometimes, SNPs occur together. A set of adjacent SNPs is termed a Multiple Nucleotide Polymorphism or MNP. Thus MNP alleles have multiple nucleotides. Microarrays for SNP detection use probes with specific targets when searching for SNPs. Next-generation sequencing (NGS) allows SNP identification without prior target information. The high coverage possible in NGS also facilitates discovery of rare alleles within population studies.

SNP or generally speaking variant detection 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 judgment is made as to the existence of a variant. Some issues that must be handled by variant detection algorithms are mentioned below:

A) **Quality of base-calls:** A sequencer outputs a sequence of nucleotides corresponding to each read. It also assigns a quality value based on the confidence with which a particular base is called. Clearly, variant detection algorithms must put greater weight on bases called with higher confidence.

B) **Mapping quality of reads:** Most alignment algorithms assign quality scores to a read based on how well the read aligned with the reference. These scores are relevant in variant detection because they measure the likelihood of a read originating from the suggested position on the reference. Even if the individual bases on a read are called with high quality values, the read may align imperfectly with the reference. The mapping quality score takes into account the insertions, deletions, and substitutions necessary for alignment at a particular position.

C) **Depth of coverage:** The number of reads covering a position also determines how confidently a variant can be called. Obviously greater sequencing depths lead to higher variant calling accuracy.

D) **Homopolymer:** A homopolymer is a repetitive sequence of a single nucleotide, e.g., AAAAAA. Sequencers often exhibit inaccurate representations of homopolymers and their immediate neighbors due to limitations in the next-generation sequencing technology. Such regions need to be handled carefully by variant detection algorithms.

E) **Ploidy:** Ploidy is the number of sets of chromosomes in a cell. Haploid organisms have one set of chromosomes per cell, while diploid organisms like humans have two. Polyploidy, the state where all cells have multiple (but more than two) sets of chromosomes is chiefly observed in plants. SNP detection must take into account the ploidy while calling SNPs. For example, in a haploid organism, each position can correspond to only one nucleotide. Reads must ideally agree at each position, and any deviation is easily detected as a sequencing error. Diploid organisms inherit one set of chromosomes from each parent, and so reads can show two nucleotides at each position, one from each set. Currently, variant analysis in Strand NGS assumes that the ploidy is two. In this case an SNP can be called under two circumstances:

- *Homozygous SNP*: One when the consensus reached from the sample reads shows a single nucleotide at the location and this consensus differs from the reference nucleotide.

- *Heterozygous SNP*: Alternatively, the sample reads may show considerable presence of two different nucleotides; typically
one of the two agrees with the reference nucleotide. The two alleles should ideally show close to 50% presence each.

In the following section, we'll describe our variant calling approach briefly and then discuss the variants detected on the whole genome sample NA12878 comprising of ~ 1.16 billion paired-end reads obtained from Illumina HiSeq 2500 sequencer [2]. These resultant variants obtained from Strand NGS will also be benchmarked with those obtained from GATK UnifiedGenotyper tool [1].

2. Our Variant Calling Approach

The variant calling algorithm in Strand NGS is capable of detecting three types of variants, namely substitution (or mutation), deletion and insertion. While reporting, these variants are categorized into four events namely substitution, insertion, deletion and complex. Substitution consists of one or more substitutions occurring in consecutive locations; similarly deletions and insertions comprise of events where one or more deletions or insertions, as the case may be, occur together. A deletion event is modeled as a mutation from a nucleotide value {A, T, G, C} to "gap". Complex events are reported when there is a mixture of multiple types of variants occurring together in consecutive locations.

The variant detection algorithm works on a per-sample basis. For each sample, data for each chromosome is analyzed, with a capability for parallel computation per chromosome. The computation is optimized for memory usage by segregating the data into windows based on maximum read size.

2.1. Selection of Potential Variant Sites

Variant detection is only performed at sites determined by pre-processing to be likely variant locations. Locations are only considered potential variant sites if they satisfy all the following requirements:

A) Read coverage threshold: The number of reads covering the location must exceed a user-defined threshold.

B) Variant coverage threshold: The number of reads covering the location and differing from the reference at the location, must also exceed a user-defined threshold.

2.2. Bayesian Algorithm for Variant Calling

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. The algorithm considers the nucleotide or the base taken by each read covering the location, as well as its associated base quality and finds the consensus genotype. The consensus genotype is the one most likely to have caused the observed data and is computed using Bayes principle.

2.2.1. Calculation of Posterior Probability

If the observed data is collectively denoted by D, then the task is to infer the consensus genotype from D. Variants are then detected by comparison to the reference sequence. The algorithm selects the genotype G that maximizes the posterior probability, $P(G \mid D)$, defined by the probability of a genotype G given the observed data. This is easily computed under Bayes principle as

$$P(G \mid D) = \frac{P(D \mid G).P(G)}{P(D)}$$

Where $P(G)$ is the prior probability of genotype G and $P(D)$ is the likelihood of data, calculated over all possible genotypes,

$$P(D) = \sum_{\forall G_i} P(D \mid G_i).P(G_i)$$

The conditional probability $P(D \mid G)$ represents the probability of observing the data D under the given genotype G.

Strand NGS specifies a score to represent the confidence with which a variant is called at a locus. This score is a function of the posterior probability of the locus being the reference genotype RR given the observed data and is defined as

$$Score = -log_{10} P(G = RR \mid D)$$

If the consensus genotype is not the reference genotype, and if the above score is greater than the user-specified cut-off, then the location is reported as a variant location and the alleles corresponding to the consensus genotype, which are different from the reference, are reported as the variant alleles.

2.2.2. Calculation of the Prior Probabilities $P(G)$

We calculate the prior probabilities of different genotypes at a given location by taking into account the following parameters.

A) The reference base at the location.

B) Heterozygosity Rate: This is representative of the prior expectation of a position being heterozygous in the organism. The default value is set to 0.001, which correlates with the population SNP rate for the human genome.

C) Homozygous to heterozygous ratio: This is the prior expectation of the ratio of the number of homozygous mutations to the number of heterozygous mutations. It has been observed that heterozygous variants are twice as frequent as homozygous variants in any diploid or human genome. Hence the default value is set to 0.5.

D) Indel to substitution ratio: This is the prior expectation of the ratio of the number of indels to the number of substitutions. Based on the information from the literature, the default value is set to 0.125.

E) Ti/Tv ratio: This is the prior expectation of the ratio of the number of transitions to the number of transversions. A transition is a point mutation that changes a purine nucleotide (A or G) to another purine, or a pyrimidine nucleotide (T or C) to another pyrimidine, whereas a transversion refers to the substitution of a purine for a pyrimidine or vice versa. Although there are twice as many possible transversions, because of the molecular mechanisms by which they are generated, transition mutations are generated at higher frequency than transversions. Expected value of this ratio for genome-wide variations is in the range 2.0 – 2.1, whereas it is in the range 3.0 – 3.3 for the exonic variations. The default value for this parameter is set to 2.6 to make it suitable for both whole genome and exome data analysis.

2.2.3. Calculation of the Conditional Probabilities $P(D \mid G)$

The computation of $P(D \mid G)$ takes into account the probability of errors observed in the data under each genotype G. It is assumed that the error in the observation of any base is independent of observation error in any other base. In most cases, error rates are derived from the associated base qualities as follows:

$$E = 10^{-(base\ quality)/10}$$

The error rates may be adjusted by taking the mapping quality into consideration. A misaligned read creates unreliable results irrespective of the base-qualities. Therefore, Strand NGS has an option to take into account the mapping qualities whenever provided by the aligner and uses the minimum of the read mapping quality and individual base quality in place of base quality for its computations.

In the absence of error rates, hard-coded error rates (that are technology-specific) are used. If error rates for a particular technology are not present in Strand NGS, it assumes a quality score of 30 for the base, which is interpreted as an error rate of 0.001.

3. Results – Benchmarking against GATK

We ran both Strand NGS and GATK variant caller on NA12878 whole-genome data using their default parameters. Strand NGS identified a total of 6,544,450 variants on the raw read set and a total of 6,335,487 variants on the processed read set which is obtained after performing steps like local realignment, base quality score recalibration and duplicate read removal. Since the focus of this article is to compare the results of Strand NGS variant caller with GATK caller, we want to remove differences that could be introduced by these pre-processing steps. Therefore, we used the GATK processed (after local realignment and base quality score recalibration) BAM file as an input in the Strand NGS variant calling algorithm as well. With this read set as an input, a summary of the variants detected by Strand NGS and GATK UnifiedGenotyper along with their characteristics is shown in table 1.

	Strand NGS	GATK
Number of total variants	6,393,054	6,105,466
Number of substitutions	5,199,005	5,085,477
Number of insertions	467,267	468,203
Number of deletions	549,043	503,689
Number of complex variants	177,739	48,097
Het/Hom ratio	2.55	2.08
Ti/Tv ratio	1.92	1.93

Table 1: Number of variants detected by Strand NGS and GATK

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%. If we only focus on substitutions, there is an overlap of ~96% between Strand NGS and GATK, confirming that for most practical purposes, variant calling results from these two algorithms are very similar.

There are a total of 444,947 variants (193,234 substitutions, 251,713 InDels and complex) identified by Strand NGS but not by GATK; and a total of 157,359 (79,706 substitutions, 77,653 InDels and complex) variants identified by GATK but not by Strand NGS.

3.1. Characteristics of Variants Identified by both Strand NGS and GATK

The hypothesis is that variants that are identified by both Strand NGS and GATK are generally of good quality. To assess this, we examined variety of characteristics of these variants. Strand NGS provides a very intuitive way to quickly assess the quality of the variants across various quality metrics like presence/absence of the variant in dbSNP database, supporting reads %, variant score, PV4 biases, etc. Figure 1 shows the histogram of supporting reads % for substitutions and 'InDels + Complex' variants. First, if we focus on substitutions, we can clearly observe a set of variants with 100% supporting reads, representing homozygous variants. For the rest, there is a peak at 50%, which is ideally expected for heterozygous variants, however there is a spread around the peak indicating the fact that heterozygous variants are not always detected at 50% supporting reads, possibly due to sequencing errors, total coverage issues, non-uniform allele sampling, etc.

Substitutions are colored by dbSNP match category with red, blue, and brown color indicating known, novel, and overlapping variants according to dbSNP138 database. In addition, it can be observed from figure 1 that most of the novel substitutions appear at the lower side of supporting read %, making them likely false positives. The distribution of supporting reads % for 'InDels + complex' has a wider spread compared to substitutions; however, a majority have a supporting read % below 50%.

Figure 2 shows the distribution of variant score that was assigned by Strand NGS to the overlapping variants (identified by both Strand NGS and GATK). Most of the variants are assigned high scores. Additionally, there is a higher fraction of novel variants than known variants on the lower side of the score, further indicating that these are likely false positives. Regardless, given the fact that larger percentage of variants are assigned a higher score, it can be concluded that most of the overlapping variants are indeed of good quality.

Figure 1:
Distribution of Supporting Read % for the variants detected by both Strand NGS and GATK (According to dbSNP137, blue indicates novel; red indicates known; brown indicates overlap)

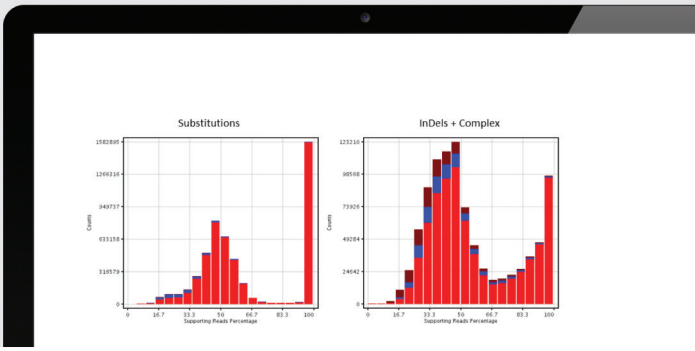

Figure 2:
Distribution of the scores assigned by Strand NGS to overlapping variants (According to dbSNP137, blue indicates novel; red indicates known; brown indicates overlap)

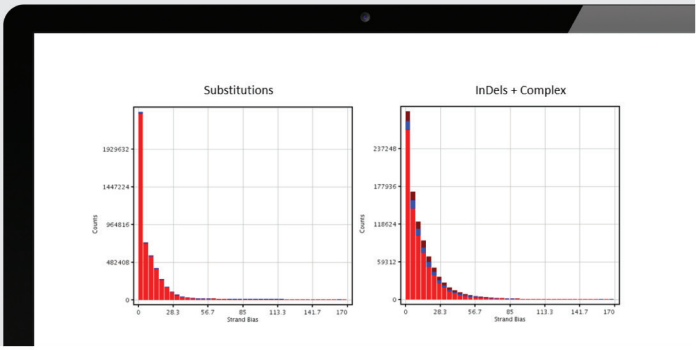

Figure 3:
Distribution of strand bias (According to dbSNP137, blue indicates novel; red indicates known; brown indicates overlap)

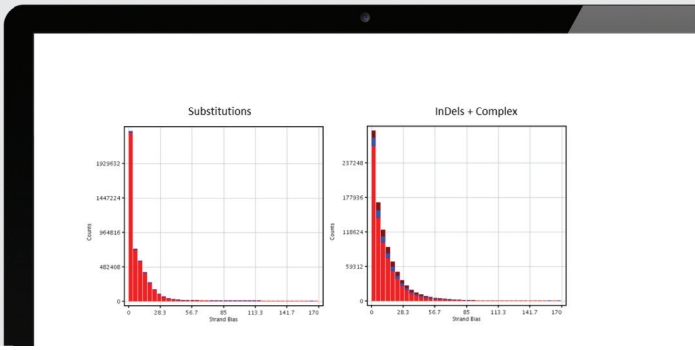

Figure 3 shows the distribution of strand bias for the same set of overlapping variants. Strand bias varies from 0 – 200, with smaller numbers indicating low strand bias. As can be seen from the figure, most of the variants have low strand bias and novel variants, and for all practical purposes are more or less uniformly distributed, again indicating overall high quality of overlapping variants.

3.2. Characteristics of Variants Uniquely Identified by Strand NGS

As mentioned, there are a total of 444,947 variants (~7%) that are identified by Strand NGS but not by GATK. Out of these, 193,234 are substitutions and 251,713 are 'InDels + complex' variants. One of the important points to note is that compared to the overlapping variants, a larger fraction of the variants in this category is novel according to dbSNP138 database.

Further, if we observe the distributions of the same metrics (as last section) for these variants, we observe that unlike the data in figure 1 and figure 2, supporting read % and variant score is skewed towards the left, as shown in figure 4 and figure 5 respectively, indicating heterozygosity of these variants with low supporting read % and low score. This also makes them likely false positives. Also strand bias as shown in figure 6 for these variants shows a substantial fraction of variants with strand bias more than 50, unlike what was observed with overlapping variants as shown in figure 3. As earlier, variants are colored by their presence or absence in dbSNP database (red: known; blue: novel; brown: overlap). These metrics therefore provide a useful as well as intuitive way to filter out the bad quality variants.

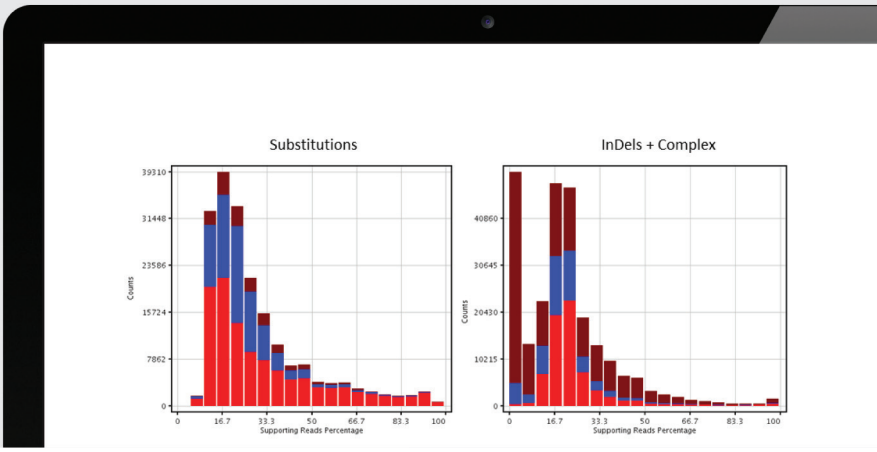

Figure 4:
Distribution of supporting read % for variants identified by Strand NGS alone

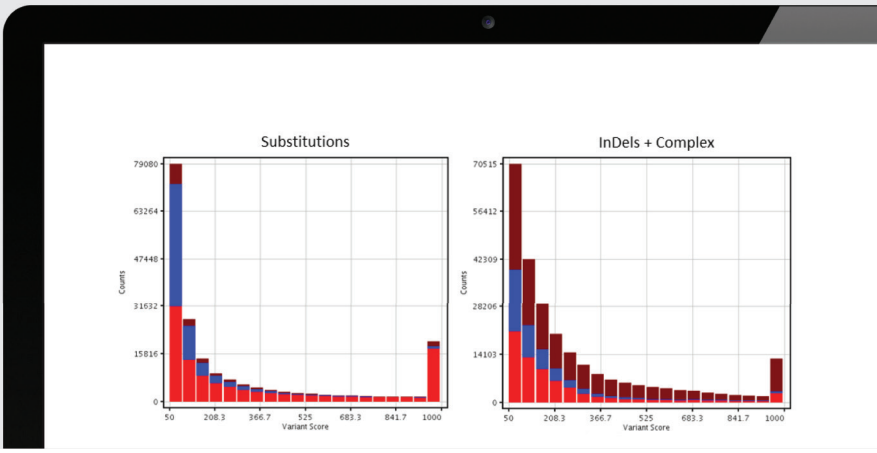

Figure 5:
Distribution of variant score for variants identified by Strand NGS alone

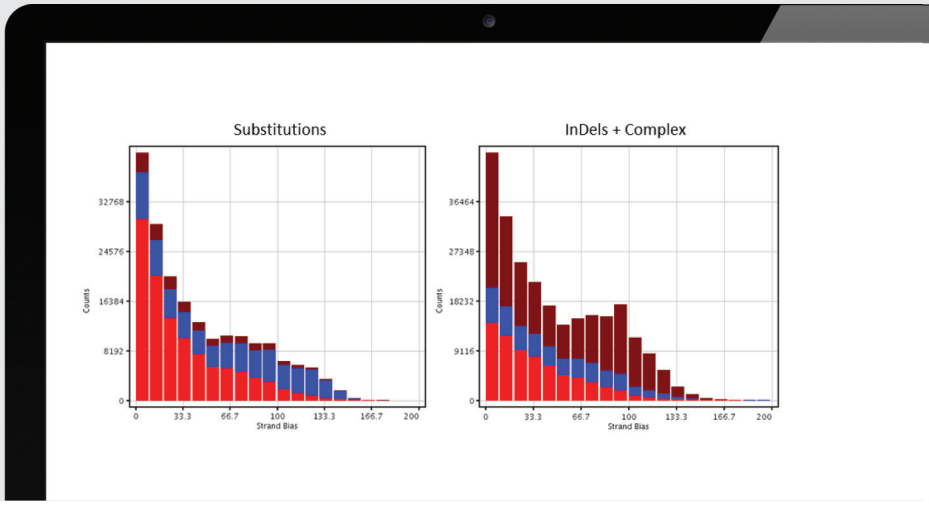

Figure 6:
Distribution of strand bias for variants identified by Strand NGS alone

Further, in addition to these metrics, Strand NGS also calculates base quality bias, mapping quality bias, tail distance bias and neighborhood quality measures. These metrics can also be used in addition to the ones used above to further filter out the bad quality variants. While Strand NGS provides a visually appealing way to examine each of these metrics individually and then filter the variants based on thresholds set on individual metrics, GATK uses a variant quality score recalibration step, which is based on a machine learning approach. The objective of this step is to use the training data and build a model describing which variant characteristics or combinations thereof makes it a bad versus good quality variant, and then recalibrate the variant scores using the model. These recalibrated scores can then be used to discard the bad quality variants. One of the downside of such a machine learning approach is the difficulty for the end-user to easily narrow down and understand a set of variables, which are actually responsible for the change in the variant quality and hence filtering. Another issue with this approach is the requirement of sufficient and good training data to build the model. One has to carefully select the training data and make sure it has the same underlying properties as the test data under consideration and further adopt ways to avoid model over-training.

As an example to show the effectiveness of the simple but visually intuitive approach of Strand NGS to filter variants using individual variables, if we set the supporting read % '>30%', variant score '>60', and strand bias '<50', we are only left with 145,324 variants in this category, making the *overlap between Strand NGS and GATK as ~98%*. This comparison makes more sense since the variant call set from GATK includes the variant quality score recalibration step.

3.3. Characteristics of Variants Uniquely Identified by GATK

There are a total of 157,359 (~2.5%) variants identified by GATK but not by Strand NGS. Of these, 79,706 are substitutions and 77,653 are 'InDels and complex'. Out of 157,359 variants in this category, ~21% are novel according to dbSNP138, again indicating these may be likely false positives. Additional evidence in favor of these variants being likely false positives may be 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, ~53% of 79,706 substitutions identified by GATK have a total coverage below 10, the default cut-off for Strand NGS.

Further, distribution of supporting read % is shown in figure 7 for variants with coverage more than or equal to 2 and with coverage more than or equal to 10. This figure clearly shows that variants with higher supporting reads % have low coverage and hence did not meet the total reads threshold of 10. Since Strand NGS uses a default of 10 reads, these variant locations are not even considered for variant calling. If we set the coverage threshold as 10, then the supporting reads % for the remaining variants is skewed towards the left, indicating many of them could be false positives and hence can be filtered out. Therefore majority of the variants uniquely identified by GATK are either not considered by Strand NGS because of low coverage or can be filtered out due to relatively bad quality.

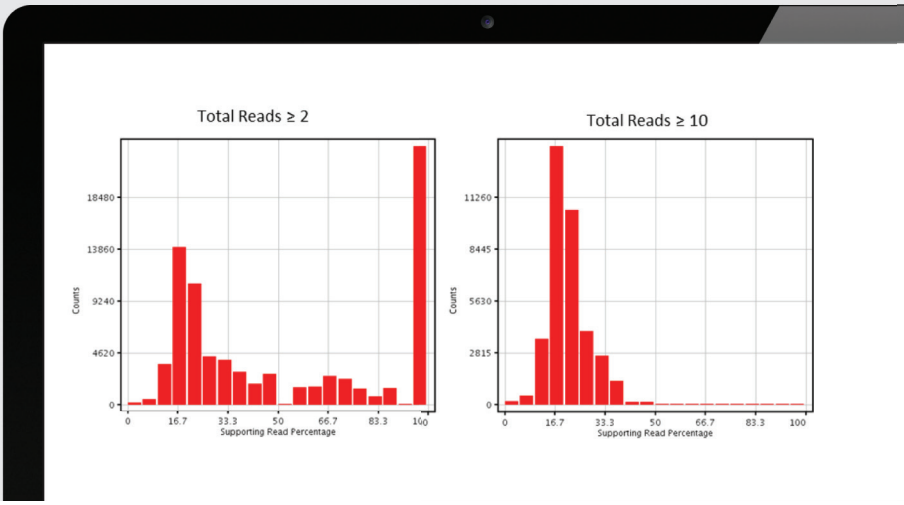

Figure 7:
Distribution of supporting reads % for variants identified by GATK alone.

4. Conclusions

In this study, we discussed the variant calling approach adopted in our software Strand NGS and presented the benchmarking results on whole-genome sample NA12878. We compared the variants identified by Strand NGS to those identified by GATK UnifiedGenotyper.

We demonstrated a high overlap (98%) in the filtered variant calls by Strand NGS and GATK. The high overlap makes Strand NGS and GATK very similar for most practical purposes. The unique variants identified by each of these algorithms first of all constitutes a very small percentage of the total variants identified and additionally many of these variants can be filtered out using quality metrics like presence/absence in dbSNP, supporting reads %, variant score, total read coverage, strand bias and other PV4 biases.

Overall, using the benchmarking results presented in this paper, we demonstrated that Strand NGS variant caller offers a solid approach for calling variants. The software also provides visually intuitive ways to intelligently filter the variant calls in order to improve specificity without affecting sensitivity. We also found the results of Strand NGS very similar to GATK, which is one of the most commonly used variant callers.

5. References

1. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D, Daly M (2011). **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nature Genetics.* 43:491-498.

2. Illumina Basespace data. [http://blog.basespace.illumina.com/2012/11/19/2x150bp-human-genome-in-record-time-with-the-hiseq-2500/]

strand

New Generation Healthcare

Strand was founded in 2000 by computer science and mathematics professors from India's prestigious Indian Institute of Science who recognized the need to automate and integrate life science data analysis through an algorithmic and computational approach. Strand's segue into the life sciences was through informatics products and services for research biologists, chemists, and toxicologists that combine advanced visualization, predictive systems modeling, data integration and scientific content management - over 2000 research laboratories worldwide (about 30% of global market share) are licensees of Strand's technology products, including leading pharmaceutical and biotechnology companies, research hospitals and academic institutions. With a recent investment by Biomark Capital, Strand has grown its established team to over 200 employees, many with multidisciplinary backgrounds that transcend computation and biology.

Since 2012, Strand has been expanding its focus to include clinical genomics, spanning sequencing, data interpretation, reporting and counseling. Strand operates a 10,000 square foot laboratory space with state-of-the-art clinical genomics capabilities and is also establishing Strand Centers for Genomics and Personalized Medicine in several hospitals around the world to serve as outreach points for genomic counseling. Based on the experience gained from sequencing, analyzing, interpreting and reporting on clinical samples over a wide variety of clinical indications, Strand has developed an end-to-end solution for clinical labs that handles all stagesfrom analysis to reporting. The interpretation and reporting software platform has been designed and developed specifically for the medical professional, ranging from the molecular pathologist to the physician. By enhancing sequence-based diagnostics and clinical genomic data interpretation using a strong foundation of computational, scientific, and medical expertise, Strand is bringing individualized medicine to the world.

For more information about Strand, please visit **www.strandls.com**,
or follow us on twitter **@StrandLife**.

INDIA
5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024

USA
548 Market Street, Suite 82804, San Francisco, CA 94104

www.strandls.com