integrity Research Journals

Journal of Animal Science and Veterinary Medicine
Volume 3(4), pages 90-93, August 2018
https://doi.org/10.31248/JASVM2018.092
ISSN 2536-7099. Copyright © 2018
Article Number: 49FBD3E01
http://www.integrityresjournals.org/journal/JASVM

Review

Variant Calling pipeline for Next Generation Sequence Data – A review

K. Ngeno

Animal Breeding and Genomics Group, Department of Animal Sciences, Egerton University, P. O. Box 536, 20115 Egerton, Kenya. Email: aarapngeno@gmail.com

Received 20th February, 2018; Accepted 30th April, 2018

ABSTRACT: Next generation sequencing (NGS) is of great significance for genetic improvement. Some of the most common application of NGS is the identification of the genomic variants, genes and sequence mutations. Mining of genomic variants such as single nucleotide polymorphisms (SNPs) from raw sequences involves several steps and use of numerous bioinformatics tools in a systematic manner. This paper reviews the components of a pipeline that calls SNPs from NGS data. The SNP calling pipeline includes base calling, quality checks, reads trimming, alignment of the quality reads to the reference genome, quality score recalibration, visualization and SNP identification. The final step of the pipeline is making biological sense out of the SNPs data, which involves filtering and annotation of the candidates SNPs.

Keywords: Annotation, Next-generation sequencing, SNP.

INTRODUCTION

Next-generation sequencing (NGS) also referred to as high-throughput sequencing, involves whole genome, specific genomic region, whole-exome or RNA sequencing. The NGS generated a lot of data which has enabled scientists to examine genomic variants such as insertions, deletions (Xi et al., 2010), single nucleotide polymorphisms (SNPs) and sequence mutations that are likely to be the causal of phenotypic variations. Single nucleotide polymorphism is the most common genomic variant that provides abundant source of genomic variation. Some of the most common application of SNPs is the studying of heritable variations (Suh and Vijg, 2005), causal genes for Mendelian diseases (Sohyun et al., 2015), mutations appropriate for diagnosis and therapy (Pabinger et al., 2014), population genetics such as phylogeography (Keim and Wagner, 2009), genome-wide associations (Nelson et al., 2014) and genomic selection. SNP-based studies rely on accurate and consistent identification of SNPs. SNP calling process from raw NGS sequences encompasses several steps and use of numerous bioinformatics tools in a systematic manner. This paper reviews the general steps and bioinformatics tools used in the pipeline that calls SNPs from NGS data and making sense out of the identified SNPs.

STEPS IN THE SNP CALLING PIPELINE

Base calling

The standard principle for NGS technologies is sequencing by synthesis. Synthesis procedure involves capturing of fluorescence images and converting into nucleotide bases to generate reads (Nielsen et al., 2011). In this process, the sequencing machines will generate errors (Altmann et al., 2012). Base calling involves estimation of sequence reads errors generated during sequencing (Nielsen et al. 2011). Estimation of the errors is usually performed by the sequencing platform using their base calling software. Errors are expressed as Phred-like quality score, which gives the expected error probability of each base call, based on the noise estimates arising from image analysis (Nielsen et al., 2011). Phred-like quality scores are calculated using the formula;

$$Q_{phred} = -10^* \log_{10} P(error)$$

where P is the probability of an incorrect base call (Pavlopoulos et al., 2013). The Phred-like scores range is 1 to 60, where a Phred score of 10, 20 and 30 represents

the accuracy of 90%, 99% and 99.9% respectively.

Quality control/trimming

Generally, all the NGS sequencing platforms do quality control checks using their base calling software and provide a summary of the data quality for the bases in the sequence reads. Although the sequencing platforms provide quality scores, there can be quality issues in sequencing. Minimizing base call errors and subsequent improvement of the accuracy of the base quality score is essential in detection of polymorphism (Nielsen et al., 2011). Therefore, quality of the generated reads needs to be checked using softwares such as prinseq (Schmieder and Edwards, 2011), shrimp2 (David et al., 2011), piqa (Martinez-Alcantara et al. 2009) or fastqc (Andrews 2010). Identified low quality read can be trimmed using a sickle (Joshi and Fass, 2011), htqc toolkit (Xi et al., 2013), solexaqa (Cox et al., 2010) or bigpre (Zhang et al., 2011).

Read mapping or Alignment

After checking quality, high-quality reads are aligned onto the reference sequences. Alignment is crucial in variant detection. Wrong alignment of the reads may result in artificial divergences from the reference sequences and consequently errors in variant calling (Nielsen et al., 2011; Altmann et al., 2012). In most cases, the accuracy of alignment relies on the alignment tool used and their corresponding settings (Altmann et al., 2012). Most common alignment tools used are either hash-based algorithms such as MAQ (Li et al., 2008b) and Stampy (Li et al. 2008a) or data compression algorithms (Burrows-Wheeler transform) like Bowtie (Langmead et al., 2009), SOAP2 (Li et al., 2009c) and BWA (Li and Durbin, 2009). These aligners generate alignments in sequence alignment map (SAM) which is converted to its binary version (BAM format). The SAM and BAM formats are the quasi-standards for storing information for the aligned sequences. The SAM and BAM files are manipulated using tools such as SAMtools (Li et al., 2009a), Genome Analysis Tool Kit (GATK) (McKenna et al. 2010) or Picard (http://picard.sourceforge.net). After the alignment step, success rate and the quality of mapped reads are usually checked by computing mapping statistics such as mean quality, quality score distribution and a fraction of reads mapped successfully. Statistics are generated using software like SAMtools and Picard. The next step is the visual inspection of the alignments. Visualization reveals the success of sequencing. Visualization is done using software tools such as GenomeView (Abeel et al., 2012), Integrative Genomics Viewer (IGV) (Robinson et al., 2011) and savant (Fiume et al., 2010).

Processing of aligned reads

Alignment post-processing step in the SNP calling pipeline involves removal of artifacts and duplicate reads using SAMtools or Picard. Presences of artifacts and duplicates will bias the SNP calling. Indels (insertions and deletions) are also identified and realigned the reads using GATK or SMRA (Homer and Nelson, 2010). Presence of indels may be mistaken as SNP (false SNP) in the subsequent analysis (Altmann et al., 2012).

Recalibration of per-base quality scores

Sequencing error rates predicted by the sequencing platforms using base-calling algorithms may not truly reflect the true base-calling errors, resulting in potential wrong SNP calls (Li et al., 2009b). Phred score values have been found to have intrinsic errors resulting in a deviation from the real sequencing errors (Nielsen et al. 2011). Therefore, there is a necessity to recalibrate the initial quality scores to improve the accuracy of the called variants (DePristo et al., 2011; Zook et al., 2012). Software used in base quality scores recalibration include SOAPsnp (Li et al., 2009b) and GATK.

Single nucleotide polymorphism calling

Single nucleotide polymorphism calling also called variant calling implies finding SNPs in the NGS data using SNP calling software. Algorithms for identifying SNPs vary in their approach. Some algorithms find SNPs based on the number of high confidence base calls that are not in agreement with the reference sequence (Olson et al., 2015). Other variant callers use likelihood ratio tests, Bayesian method or machine learning statistical methods that consider allele frequencies, base and quality scores to call SNPs (Nielsen et al., 2011; Pabinger et al., 2014). The most commonly used SNP callers are SAMtools, GATK, Freebayes (Garrison and Marth, 2012), VarScan (Koboldt et al*.,* 2009), cortex_var (Iqbal et al., 2012), VCFtools (Danecek et al., 2011) and Torrent Variant Caller (https://www.thermofisher.com). The end result of SNP calling is the collection of SNPs in a standard file called Variant Call Format (VCF) which is generated by SNP callers.

Filtering and annotation of SNP candidates

In the SNP identification step, the posterior probabilities calculated from each site may deviate from the true value due to errors (Nielsen et al., 2011). Therefore, additional filtering is applied to improve SNP calls by removing false positive SNPs and SNP calling artifacts. Filtering can be based on the posterior probabilities, read depth, quality scores differences between major and minor alleles, linkage disequilibrium patterns, strand biases and deviation from Hardy–Weinberg equilibrium (Nielsen et al., 2011; Altmann et al., 2012). VCFtools, GATK and

SAMtools are tools commonly used in filtering SNPs. The final step of the pipeline is making biological sense out of the called SNPs, which involves carrying out annotation to predict their potential effects or functions. Annotation of SNPs involves extraction of biological information base on nucleic acid and protein sequence. Commonly used annotation tools are SNPeff (Cingolani et al., 2012), VEP-Variant Effect Predictor (McLaren et al., 2010), ANNOVAR (Wang et al., 2010) , PolyPhen-2 (Adzhubei et al., 2010), SIFT (Ng and Henikoff, 2003) and FAST-SNP (Saa and Nielsen, 2016).

CONCLUSION

This paper provides useful guidelines for reliable SNP calling from NGS data to the annotation of the identified SNPs. SNP calling is a multistep process involving several bioinformatics tools. As such, SNP calling has to be carried out using steps, methods and tools that have been tested and benchmarked.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest

REFERENCES

Abeel, T., Van, P. T., Saeys, Y., Galagan, J., & Van d. P. Y. (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Res,* 40, 12.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7, 248.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., & Müller-Myhsok, B. (2012). A beginner's guide to SNP calling from high-throughput DNA-sequencing data. *Human genetics,* 131, 1541-54.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2), 80-92.

Cox, M., Peterson, D., & Biggs, P. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, 11, 485.

David, M., Dzamba, M., Lister, D., Ilie, L., & Brudno, M. (2011). SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics*, 27, 1011-1012.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis A. A., Del, Angel, G., Rivas, M. A., & Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43, 491-8.

Fiume, M., Williams, V., Brook, A., & Brudno, M. (2010). Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, 26, 1938-44.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

Homer, N., & Nelson, S. (2010). Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol*., 11, 99.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44, 226.

Joshi, N., & Fass J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Keim, P. S., & Wagner, D. M. (2009). Humans and evolutionary and ecological forces shaped the phylogeography of recently emerged diseases. *Nat. Rev. Microbiol*., 7, 813-821.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., & Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25, 2283-2285.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*.,10, 25.

Li H., & Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin R. (2009a). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25, 2078-9.

Li, H., Ruan, J. & Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using.

Li, H., Ruan, J. & Durbin, R. (2008b). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18, 1851-1858.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., & Kristianse, K. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res*., 19, 1124-1132.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., & Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19, 1124-1132.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., & Wang, J. (2009c) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.

Martinez-Alcantara, A., Ballesteros, E., Feng, C., Rojas, M., Koshinsky, H., Fofanov, V., Havlak, P., & Fofanov, Y. (2009). PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. Bioinformatics 25 (18), 2438-2439.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20, 1297-1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26, 2069-2070.

Nelson, C. L., Pelak, K., Podgoreanu, M. V., Ahn, S. H., Scott, W. K., & Allen, A. S. (2014). A genome-wide association study of variants associated with acquisition of Staphylococcus aureus bacteremiaina healthcare setting. *BMC Infect. Dis*., 18, 83.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011) Genotype and SNP calling from next-generation sequencing

data. *Nature Reviews Genetics*, 12, 443-451.

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31, 3812-3814.

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow J. B., Salit, M. L., & Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Frontiers in genetics, 6, 235.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics,* 15, 256-78.

Pavlopoulos, G., Oulas, A., Iacucci, E., Sifrim, A., Moreau, Y., & Schneider, R. (2013). Unraveling genomic variation from next generation sequencing data. *Bio Data Min*. 6(1), 13.

Robinson, J., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E., Getz, G., & Mesirov, J. (2011). Integrative genomics viewer. *Nat. Biotechnol*., 29, 24-26.

Saa, P. A., & Nielsen, L. K. (2016). Fast-SNP: a fast matrix pre-processing algorithm for efficient loopless flux optimization of metabolic models. *Bioinformatics*, 32, 3807-3814.

Schmieder, R., & Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27, 863-864.

Sohyun, H., Eiru, K., Insuk, L., & Edward, M. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Nature*, 5, 17875.

Suh, Y., & Vijg, J. (2005). SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 573, 41-53.

Torrent Variant Caller: Torrent Suite™ Software. Available at https://www.thermofisher.com

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., 38, 164.

Xi, R., Kim, T., & Park, P. (2010). Detecting structural variations in the human genome using next generation sequencing. *Brief Funct Genomics*, 9, 405-415.

Xi, Y., Di, L., Fei, L., Jun, W., Jing, Z., Xue, X., Fangqing, Z., & Baoli, Z. (2013). HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics,* 14, 33.

Zhang, T., Luo, Y., Liu K., Pan, L., Zhang, B., Yu, J., & Hu, S. (2011). BIGpre: a quality assessment package for next-generation sequencing data. *Genomics Proteomics Bioinformatics*, 9, 238-244.

Zook, J. M., Samarov, D., McDaniel, J., Sen, S. K., & Salit, M. (2012) Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *PloS one* 7, e41356.