3 Biotech (2016) 6:68
DOI 10.1007/s13205-016-0382-1

ORIGINAL ARTICLE

VDAP-GUI: a user-friendly pipeline for variant discovery and annotation of raw next-generation sequencing data

Ramesh Menon[1] · **Namrata V. Patel**[1] · **Amitbikram Mohapatra**[1] · **Chaitanya G. Joshi**[1]

Received: 19 June 2015 / Accepted: 15 October 2015 / Published online: 15 February 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Even though next-generation sequencing (NGS) has become an invaluable tool in molecular biology, several laboratories with NGS facilities lack trained Bioinformaticians for data analysis. Here, focusing on the variant detection application of NGS analysis, we have developed a fully automated pipeline, namely Variant Discovery and Annotation Tool-Graphical User Interface (VDAP-GUI), which detects and annotates single nucleotide polymorphisms and insertions/deletions from raw sequence reads. VDAP-GUI consolidates several proven methods in each step such as quality control, trimming, mapping, variant detection and annotation. It supports multiple NGS platforms and has four methodological choices for variant detection. Further, it can re-analyze existing data with alternate thresholds and generates easily interpretable reports in html and tab-delimited formats. Using VDAP-GUI, we have analyzed a publically available human whole-exome sequence dataset. VDAP-GUI is developed using Perl/Tk programming, and is available for free download and use at http://sourceforge.net/projects/vdapgui/.

Keywords High-throughput sequencing · Single nucleotide polymorphism · INDEL mutation · Software pipeline

Introduction

In the last decade the massively parallel next-generation sequencing (NGS) technology has revolutionized life science research, which pave the way for several landmark discoveries (Metzker 2010; Shendure and Ji 2008). For example, a single experiment can identify thousands of single nucleotide variants (SNVs) and small INDELs associated to a genetic disorder (Stitziel et al. 2011). With the introduction of affordable desktop NGS platforms several small and medium sized laboratories use this technology for whole genome/exome, transcriptome or metagenome applications (Gullapalli et al. 2012). Even though some NGS data analysis pipelines in the free domain are frequently updated to handle the huge datasets, more efficient and user-friendly analytical solution need to be evolved, especially in variant discovery applications. Present challenges in NGS data analysis includes computation power, storage, data sharing, expensive commercial softwares and requirement of skilled professionals (Sarovich and Price 2014). This apart, the fragmented nature of NGS softwares makes a comprehensive analysis out of reach for many life science researchers. Although few automated pipelines have been developed recently, many of them lack a user-friendly interface or are limited by functionality (Blanca et al. 2011; D'Antonio et al. 2013; Sarovich and Price 2014).

Here, we have developed VDAP-GUI, an open-source tool for variant detection in eukaryotic genome/exome data. VDAP-GUI wraps several well-validated open-source

Ramesh Menon, Namrata V. Patel and Amitbikram Mohapatra have contributed equally.

Electronic supplementary material The online version of this article (doi:10.1007/s13205-016-0382-1) contains supplementary material, which is available to authorized users.

✉ Chaitanya G. Joshi
cgjoshi@aau.in;
http://www.omeresearchfacility.in

[1] Department of Animal Biotechnology, College of Veterinary Sciences and Animal Husbandry, Anand Agricultural University, Anand, Gujarat 388 001, India

programs into a single-platform, thereby simplifying and standardizing the analysis workflow. The methods in VDAP-GUI include FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality control, PRINSEQ (prinseq.sourceforge.net/) for trimming, Burrows-Wheeler Aligner (BWA) (Li and Durbin 2010) for reference mapping. For SNP/INDEL detection it offers three standard methods such as SAMtools (Li et al. 2009), VarScan (Koboldt et al. 2009) and FreeBayes (Garrison and Marth 2012). Further, the variant annotation is carried out using variant effect predictor—VEP tool (http://www.ensembl.org/info/docs/tools/vep), which supports annotation of 65 eukaryotes.

Materials and methods

Tools used in VDAP-GUI and development of pipeline

The methods in VDAP-GUI include FastQC—version 0.11.2 (www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality control and PRINSEQ—version 0.20.4 (prinseq.sourceforge.net/) for quality filter and trimming. For reference mapping, BWA-mem (version 0.7.5a) algorithm has been utilized (Li and Durbin 2010). The SNP/INDEL detection methods used in VDAP-GUI were: SAMtools (version 0.1.19), VarScan (version 2.3.7) and FreeBayes (version 0.9.10-3) (Garrison and Marth 2012; Koboldt et al. 2009; Li et al. 2009). Further, a custom approach namely, *MultiCom* is also introduced to detect SNPs detected by more than one method. The *VCF-tools intersect* method has been used to derive *MultiCom* results. For variant annotation, Ensembl VEP tool version 78 (http://www.ensembl.org/info/docs/tools/vep) was employed, which supports annotation of 65 eukaryotes.

Initially, the tools used in VDAP-GUI were compiled on Ubuntu Linux (version 14.04) and integrated into a pipeline using Perl language (version 5.20), followed by testing it in the command line version. Next, the code was subjected to necessary modification after addition of Tk module (version 804.032) for GUI mode. In the pipeline, some in-house Perl/bash scripts were used for the automation process.

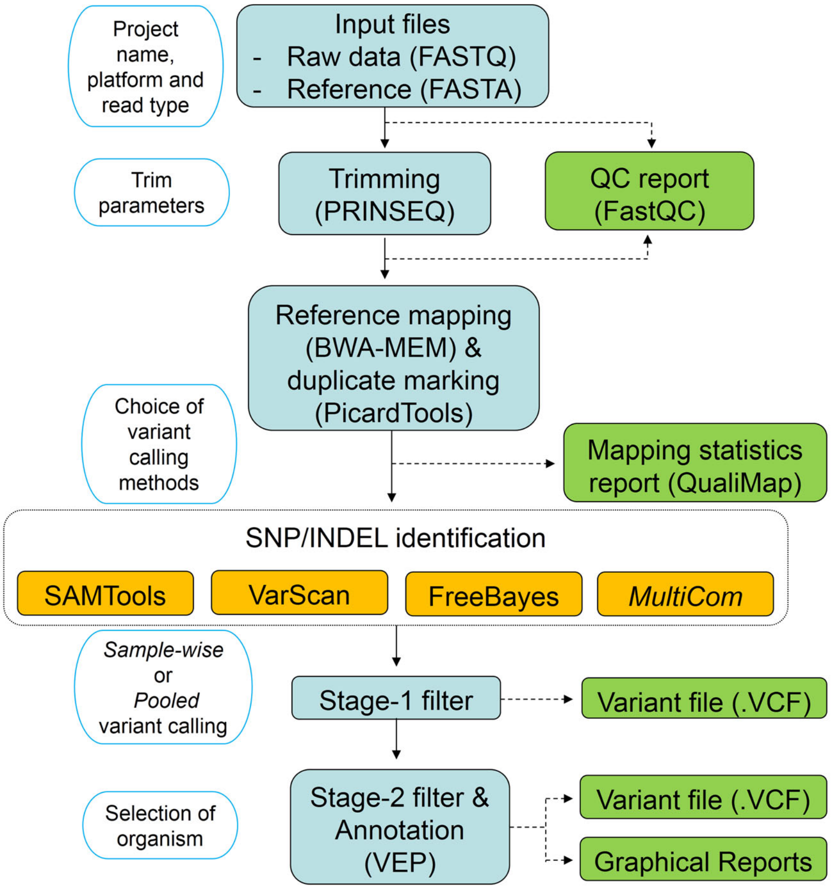
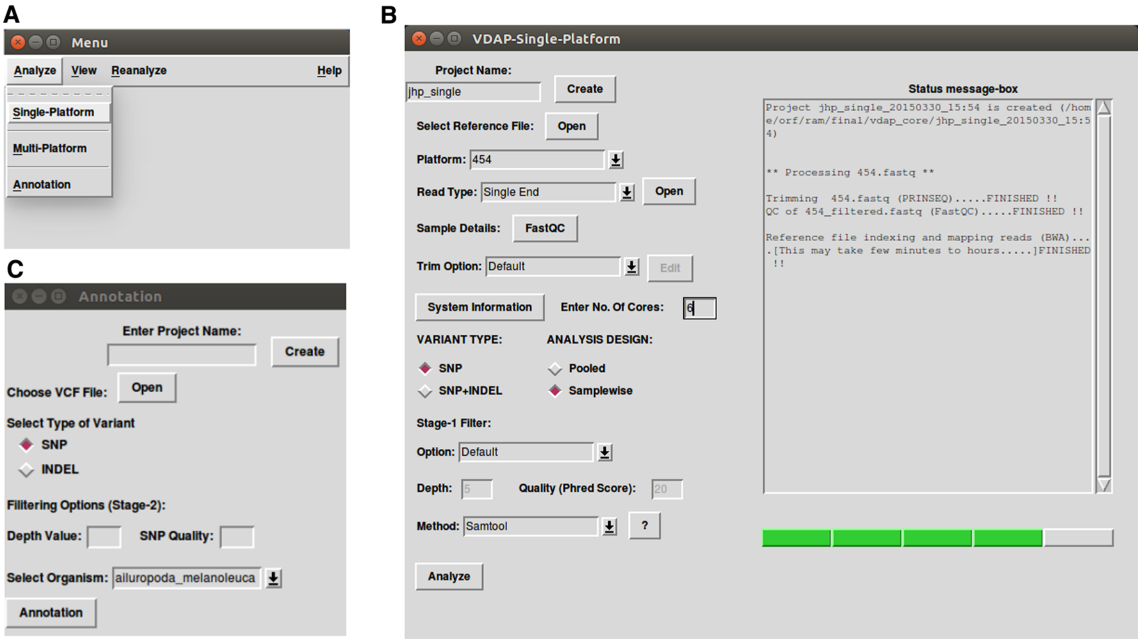


Fig. 1 (*Workflow of VDAP-GUI*) the *left section* shows the user input to the software. The *middle section* indicates the processes and softwares used. In the variant calling process four options are possible *S* SAMtools, *V* VarScan, *F* FreeBayes and *M* Multicom. The *right section* shows the outputs given at each step by VDAP-GUI. The *arrows* indicate dependency between each step. The *upper panel* shows the analysis pipeline and the *lower panel* shows the variant annotation module



Project Name: jhp_single

Platform: 454

Read Type: Single End

Sample Details:

Trim Option: Default

Enter No. Of Cores: 6

VARIANT TYPE:
- ● SNP
- ○ SNP+INDEL

ANALYSIS DESIGN:
- ○ Pooled
- ● Samplewise

Stage-1 Filter:

Option: Default

Depth: 5 Quality (Phred Score): 20

Method: Samtool

Status message-box

```
Project jhp_single_20150330_15:54 is created (/hom
e/orf/ram/final/vdap_core/jhp_single_20150330_15:5
4)


** Processing 454.fastq **

Trimming  454.fastq (PRINSEQ).....FINISHED !!
QC of 454_filtered.fastq (FastQC).....FINISHED !!

Reference file indexing and mapping reads (BWA)...
.[This may take few minutes to hours.....]FINISHED
 !!
```

Select Type of Variant
- ● SNP
- ○ INDEL

Select Organism: ailuropoda_melanoleuca

Fig. 2 (*Screenshots of VDAP-GUI*) the *panel A* shows the main Menu of VDAP-GUI. The *Analyze* menu has items such as *Single-platforms*, *Multi-platform* and *Annotation*. The *View* menu has *Results* item. The Re-*analyze* menu has items such as *Single-platforms* and *Multi-platform*. The *Help* has *Manual*, *Credits*, and *About* items. The *panel B* shows the *Analyze* → *Single-platform* module, in which the *left* side has the input and parameter setting and *right* hand side has a status message-box, to display the real-time status of the analysis along with a progress *bar*. In the *Panel C*, *Annotation* module is given. User is prompted to provide the project name and choose the .vcf file along with the Stage 2 filter (optional). In addition, the user is provided with a drop-down list for the selection of organisms (currently, 65 eukaryotes are supported for annotation)

Results and discussion

Overview of VDAP-GUI workflow

The VDAP-GUI tool wraps a collection of publicly available NGS analysis tools using Perl/Tk programming. The workflow of VDAP-GUI pipeline is depicted in Fig. 1, which consists of five steps: (1) *create project/import input files*, (2) *QC and Trimming*, (3) *reference mapping*, (4) *SNP/INDEL identification*, (5) *annotation*. Further, this menu-driven software is highly customizable at different steps of analysis.

Creating project and input file import

The analysis of raw sequence data can be evoked by selecting the *Analyze* item from the main menu (Fig. 2a). The pipeline needs to be initialized with a project creation, followed by selection of raw read dataset (FASTQ format) and a reference sequence file in FASTA format. In addition, appropriate platform name and read type (single or paired-end) need to be selected from the drop-down list (Fig. 2a).

QC and trimming

In the QC analysis, the VDAP-GUI makes assessments about the quality of the reads, read length, GC content, k-mer analysis using FASTQC tool. Though optional, QC analysis of the raw reads is recommended to understand the overall quality of the dataset and estimate the parameters for filtering of low quality reads. However, QC will be performed in background of the filtered dataset. In both the cases, detailed reports will be stored in the project folder (Fig. 2b). The filtering of raw reads is performed using PRINSEQ tool, in which the user can choose between *default* and *custom* option. The *default* trim options are specific for each NGS platform (Fig. 2b). However, the filtering of raw reads can be skipped by choosing the appropriate parameters.

Reference mapping and duplicate marking

The reliable reads are mapped to the reference genome in this step using bwa-mem algorithm. Initially, the workflow checks whether the reference sequence is indexed or not. If not, indexing will be performed using tabix tool. After reference mapping, the alignment file will be subjected to duplicate marking using Picard tools. The resultant SAM file will be converted to BAM format and in the project folder. In general, reference mapping and duplicate marking are the most time consuming step in VDAP-GUI.

SNP/INDEL identification

The choice of the variant calling algorithm is one important factor that determines the reliability of the SNP/INDEL. VDAP-GUI provides four choices for SNP discovery, among which three are widely used algorithms and one is a strategy. The algorithms are SAMtools, VarScan and Freebayes. The strategy is named as *MultiCom*, which performs the variant discovery using all the three algorithms and selects those which are identified by at least two algorithms. At present, *MultiCom* feature is not implemented for INDEL discovery. VDAP-GUI reports variants based on the base depth and quality parameters given by the user, named as *Stage-1 filter*. The resultant file (.vcf) will be saved in the project folder.

Annotation

Variant annotation is the final step in the VDAP-GUI. This step is implemented as an independent module, as this feature can also be utilized for annotation of variants obtained by tools other than VDAP-GUI (Fig. 2c). In this step, the user may opt for a second level of filtering (namely *Stage-2 filter*), to select the most reliable variants for annotation. The annotation is performed using VEP tool, which supports 65 eukaryotes. Apart from the report generated by VEP tool, VDAP-GUI provides detailed reports on variant prediction in html and csv formats.

Additional features

VDAP-GUI offers the *re-analysis* feature, which will be useful if the user needs to analyze the same dataset repeatedly with varying parameters and algorithms. In this case, the user can choose the project folder and the settings are automatically listed in the input form. After necessary changes in the parameter setting, the user can re-analyze the dataset and the result will be stored as a sub-folder in the main project folder. In addition, VDAP-GUI offers a comprehensive results viewing window, in which the user can view all the results of the project without navigating away from VDAP-GUI. Lastly, this tool can be executed in multiple instances at the same time.

Comparison with other tools

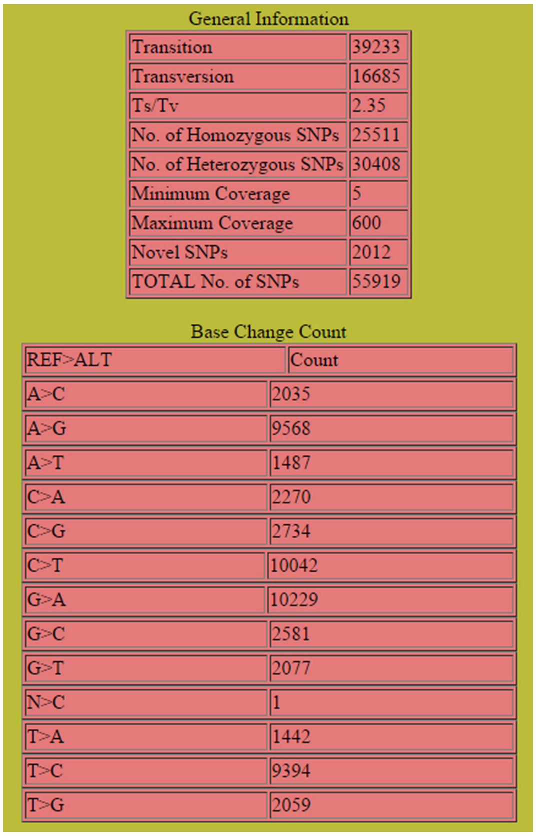
In the recent years, variant detection and annotation tools have been developed in the free as well as commercial domains (Table 1). SPANDx is one such pipeline which was developed using Linux bash script (Sarovich and Price 2014). This pipeline works similar to VDAP-GUI, and the authors have incorporated phylogenetic analysis too. However, this command line pipeline does not take care of the quality control/trimming of the raw sequences (Table 1). Another tool namely WEP, is a web-based software providing graphical interfaces for variant discovery from raw NGS data (D'Antonio et al. 2013). Practically, uploading of huge datasets generated by the high-throughput experiments like NGS is not convenient. In addition, the access of the WEP tool is restricted, the analysis is specific for or limited to human exome sequences generated only by the Illumina platform. More recently, fastq2vcf tool has been developed, which has features very close to VDAP-GUI (Gao et al. 2015). This PERL based pipeline supports multiple platforms and single/paired-end reads. However, fastq2vcf is a command line tool, and the user needs to edit the configuration file before starting any analysis. This apart, fastq2vcf does not provide the option for trimming of raw sequences. Similarly, the ngs-backbone tool, a command line driven tool (Blanca et al. 2011), is restricted to the analysis of transcriptome data. Differently, the SeqGene tool is designed to handle both transcriptome and whole-exome datasets (Deng 2011). However, this command line tool does not provide options for trimming of raw reads. GAMES stand-alone tool is capable of handling whole genome as well as whole-exome datasets and generates excellent reports (Sana et al. 2011). However, this tool has command line interface and does not give options to process raw data. Finally, we have compared VDAP-GUI features with CLCBio genomics workbench 7.0.3 (http://www.clcbio.com). This commercial software has excellent graphical interface compared to VDAP-GUI and is much more user-friendly and intuitive, compared to similar tools in free domain. The software uses a proprietary algorithm for variant detection. It is also capable of detecting copy number variation (CNVs). However, the highly expensive licensing is a major limiting factor for this tool. Finally, we have excluded few popular pipelines (e.g., GATK, Atlas2) from the comparison with VDAP-GUI, as they do not process .fastq files, but require aligned files in SAM/BAM format in the first step (DePristo et al. 2011; Evani et al. 2012).

Integration of VEP annotation tool enables VDAP-GUI to annotate the variants from more than 65 eukaryotes.

The other widely used variant annotation tools are SnpEff and ANNOVAR (Cingolani et al. 2012; Wang et al. 2010). In VEP, protein functional consequence prediction

Table 1 Feature comparison of variant detection and annotation pipelines

Features	VDAP-GUI	SPANDx	WEP	fastq2vcf	ngs_backbone	SeqGene	GAMES	CLCBio
Web-based/Stand-alone	Stand-alone	Stand-alone	Web-based	Stand-alone	Stand-alone	Stand-alone	Stand-alone	Web-based/stand-alone
Free for academic use	Yes	Yes	Yes (restricted access)	Yes	Yes	Yes	Yes	No
GUI	Yes	No	Yes	No	No	No	No	Yes
Organisms	No restriction	No restriction	Human	No restriction	No restriction	No restriction	No restriction	No restriction
Platforms supported	Illumina/454/ion torrent	Illumina/454/ion torrent	Illumina	Multiple	Multiple	Multiple	Multiple	Illumina/454/ion torrent
Mixed platform support	Yes	No	No	Yes	No	No	No	Yes
Paired-end support (PE)	Yes (PE for Illumina)	Yes (PE for Illumina)	Yes	Yes	Yes	Yes	Yes	Yes
Quality control	Yes	No	Yes	Yes	Yes	Yes	No	Yes
Sequence trimming	Yes (Default/Custom)	No	Yes	No	Yes	No	No	Yes (Default/Custom)
WGS	Yes	Yes	No	Yes	No	No	Yes	Yes
WES	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
RNA-seq	No	No	No	No	Yes	Yes	No	Yes
Tools for variant call	SAMtools/VarScan/Freebayes	SAMtools/GATK	GATK	SAMtools/GATK/SNVer	SAMtools/GATK	SeqGene SNP pileup	GAMES method	Proprietry method
Variant types	SNP/INDEL	SNP/INDEL	SNP/INDEL	SNP/INDEL	SNP/INDEL	SNP/INDEL/CNV	SNP/INDEL/CNV	SNP/INDEL/CNV
Annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes, limited
Reports	Yes	No	Yes	Yes (limited)	Yes (limited)	Yes (limited)	Yes	Yes



General Information

Transition	39233
Transversion	16685
Ts/Tv	2.35
No. of Homozygous SNPs	25511
No. of Heterozygous SNPs	30408
Minimum Coverage	5
Maximum Coverage	600
Novel SNPs	2012
TOTAL No. of SNPs	55919

Base Change Count

REF>ALT	Count
A>C	2035
A>G	9568
A>T	1487
C>A	2270
C>G	2734
C>T	10042
G>A	10229
G>C	2581
G>T	2077
N>C	1
T>A	1442
T>C	9394
T>G	2059

Fig. 3 (*Summary of SNPs discovered in human whole-exome dataset*) the *upper panel* shows the general information of the Stage 1 detected SNPs such as Ts/Tv ratio, zygosity, number of SNVs, etc. The *lower panel* gives the frequency of SNPs in terms of nucleic acid changes. *REF* and *ALT* refers to bases in reference and sample, respectively, and *Count* indicates the frequency of each base change

is available for 10 organisms, including human, rat, cow, dog, etc.

VDAP-GUI is a unique combination of user-friendliness, customization and functionality. It can assist non-IT expert researchers who do not prefer to work with Linux command line. Notably, we have tested VDAP-GUI in Microsoft Windows 7 platform running with Virtual box software (https://www.virtualbox.org/).

Test dataset and assessment of the pipeline

We have considered a publicly available human whole-exome sequence dataset namely PDA_033-Tumor (NCBI SRA ID: SRX968817), which is a part of study titled "*Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets*" (Witkiewicz et al. 2015). The sequencing experiment had been performed using Illumina HiSeq2500 platform. The raw fastq files (paired-end) have been downloaded from NCBI SRA database to VDAP-GUI software along with the Human genome reference genome version hg19 (.fa file). In the workflow, we performed quality control of raw datasets (FastQC) and followed the default parameters for trimming (PRINSEQ) of Illumina paired-end NGS sequence dataset, i.e., phred quality score ≥20 and length ≥40 bp. The average coverage of the dataset was found to be about 36×. In VDAP-GUI, we chose variant type as 'SNP' and *analysis design*, with the default Stage-1 filter: depth = 5, quality = 20. SAMtools method was chosen for SNP discovery. With the given parameters, the pipeline discovered a total of 55,919 SNPs (homozygous SNPs = 25,511, heterozygous SNPs = 30,408) detected with transition-transversion (ts/tv) ratio of 2.35 (Fig. 3). Next, we loaded the resultant *.vcf* file to the *annotation* module of VDAP-GUI with Stage-2 filter criteria of depth = 10, quality = 30, to annotate the high-confidence SNPs, which resulted in the annotation of 46,963 SNPs (homozygous SNPs = 19,252, heterozygous SNPs = 27,712). Interestingly, out of 46,963 SNPs, 1537 have not been reported previously (Supplementary file 1 and 2). In an Ubuntu Linux-based desktop equipped with Intel i7 quad-core processor (8 MB L3 cache memory) and 16 GB RAM, the SNP discovery pipeline finished in 2 h and 44 min, and the annotation analysis took about 1 1/2 h till the generation of all annotation output files.

Conclusions

VDAP-GUI is a simple and robust open-source tool developed primarily for non-IT expert researchers. It incorporates several proven tools at each step of variant discovery pipeline. Apart from the essential elements for variant discovery, VDAP-GUI's unique features make it distinct in the non-commercial domain of variant discovery pipelines.

Acknowledgments The study was supported by Department of Biotechnology, Ministry of science and technology, India.

Authors' contributions RM and CGJ conceived and designed the study. NVP, AM and RM developed the software. NVP and RM analyzed the data. RM supervised the study. RM and CGJ wrote the manuscript. All authors have read and approved the final version of the manuscript.

Compliance with ethical standards

Conflict of interest There are no conflict interests.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Blanca JM, Pascual L, Ziarsolo P, Nuez F, Canizares J (2011) ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. BMC Genom 12:285. doi:10.1186/1471-2164-12-285

Cingolani P et al (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 6:80–92. doi:10.4161/fly.19695

D'Antonio M et al (2013) WEP: a high-performance analysis pipeline for whole-exome data. BMC Bioinform 14(Suppl 7):S11. doi:10.1186/1471-2105-14-S7-S11

Deng X (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome- sequencing data. BMC Bioinform 12:267. doi:10.1186/1471-2105-12-267

DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. doi:10.1038/ng.806

Evani US et al (2012) Atlas2 Cloud: a framework for personal genome analysis in the cloud. BMC Genom 13(Suppl 6):S19. doi:10.1186/1471-2164-13-S6-S19

Gao X, Xu J, Starmer J (2015) Fastq2vcf: a concise and transparent pipeline for whole-exome sequencing data analyses. BMC Res Notes 8:72. doi:10.1186/s13104-015-1027-x

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

Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ (2012) Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. J Pathol Inform 3:40. doi:10.4103/2153-3539.103013

Koboldt DC et al (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285. doi:10.1093/bioinformatics/btp373

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26:589–595. doi:10.1093/bioinformatics/btp698

Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352

Metzker ML (2010) Sequencing technologies: the next generation. Nat Rev Genet 11:31–46. doi:10.1038/nrg2626

Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S (2011) GAMES identifies and annotates mutations in next-generation sequencing projects. Bioinformatics 27:9–13. doi:10.1093/bioinformatics/btq603

Sarovich DS, Price EP (2014) SPANDx: a genomics pipeline for comparative analysis of large haploid whole genome re-sequencing datasets. BMC Res Notes 7:618. doi:10.1186/1756-0500-7-618

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145. doi:10.1038/nbt1486

Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12:227. doi:10.1186/gb-2011-12-9-227

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164. doi:10.1093/nar/gkq603

Witkiewicz AK et al (2015) Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nat Commun 6:6744. doi:10.1038/ncomms7744