Full Length Research Article
Advancements in Life Sciences – International Quarterly Journal of Biological Sciences

ARTICLE INFO

Date Received: 18/05/2019;
Date Revised: 16/05/2020;
Date Published Online: 25/05/2020;

Authors' Affiliation:
1. Institutes of Biotechnology, Gulab Devi Educational Complex, Lahore - Pakistan
2. Decode Genomics, 323-D, Punjab University Employees Housing Scheme (Town II), Lahore - Pakistan
3. Institutes of Biotechnology, Gulab Devi Educational Complex, Lahore – Pakistan
4. Department of Biology, Tabuk University - Kingdom of Saudi Arabia

***Corresponding Author:**
Rashid Saif
Email:
rashid.saif37@gmail.com

How to Cite:
Saif R, Ejaz A, Mehmood T, Asif F, Alghanem SA, Ahmad TS (2020). Introduction to Galaxy Platform for NGS Variant Calling Pipeline. Adv. Life Sci. 7(3): 129-134.

Keywords:
Galaxy platform; NGS data; Teddy goat; Variant calling; Bioinformatics

Open Access

Introduction to Galaxy Platform for NGS Variant Calling Pipeline

Rashid Saif*[1,2], Aniqa Ejaz[3], Tania Mehmood[3], Fatima Asif[3], Suliman Mohammad Alghanem[4], Talha Saleem Ahmad[2]

Abstract

Background: Galaxy web-based platform for Next Generation Sequence (NGS) data analysis provides unprecedented opportunities to characterize, analyze and computationally visualize genomic landscapes with limited-resources. An initiative was taken to explore this pipeline for NGS data-analysis by using Galaxy platform, for its relative accessibility, reproducibility, transparency and scalability.

Methods: Variant calling and associated workflows were executed on NGS pooled-seq data of 12 Pakistani Teddy goats. Different tools used in this pipeline are FastQC for quality checks, Trimmomatic for trimming data, SAM/BAM tools for conversion of file formats, Picard tools for marking deduplicates, VCFtools/FreeBayes for genomic variant detection and SnpSift to annotate the variants.

Results: Highly associated functionally untrivial 43,712 loci were percolated having 87,510 alleles. Besides, 1,548 variants with 1,134 SNPs, 23 mixed variants, 76 MNP, 183 insertions and 132 deletions were observed in Teddy breed using San Clement ARS1 reference genome. Furthermore, 1,283 homozygous and 265 heterozygous variant were also divulged out of 43,447 loci. These variants are likely to be liable for general phenotypic traits of Teddy with smaller body-size, tender meat quality and agility along with other breed specific traits.

Conclusion: Galaxy fulfills the core function of reproducibility and easy accessibility by removing the gaps between large data analysis and its interpretations. This variant calling pipeline reveals the genomic differences of Teddy specific characteristics as compare to ARS1 reference genome.

SCAN ME

Introduction

Genomic data science posed significant challenges to the scientific community, which led to the development of plethora of Next Generation Sequencing (NGS) tools for providing solution by introducing systems, platforms, scripting languages and primary & secondary databases [1]. Galaxy platform is one of the examples which is equipped with various powerful tools used in bioinformatics with free accessibility for genomics, metagenomics, biostatistics and comparative genomics data analyses [2]. In order to test various bioinformatics tools for variant calling analysis, we did hands-on estimations and took whole genome pooled-sequence data of Teddy goat which was subjected to variant calling pipeline in this study to identify potential breed specific variants.

Teddy is one of the most important breed of goats with significant demand in Pakistan due to its small body-size, low fodder intake, high meat production and competitive breeding potential. The home tract of Teddy goat in Pakistan is Southern Punjab and Azad Jammu & Kashmir [3]. They have a reasonable population size with weaning and yearling weight of 10 kg and 20 kg respectively [4]. In this whole genome comparison study, Teddy goat was analyzed against San Clemente to look into their breed specific traits. It is anticipated that the increased demand of Teddy goat is due to its ability of multiple births, meat production/quality and its ability to thrive in harsh weather conditions of Pakistan [5], while reference genome ARS1 is San Clemente breed from California-USA taken from NCBI is a robust, medium-short stature, multipurpose breed having good potential to serve as a reference [6].

Achievements in bioinformatics software and tools provide numerous applications for analyzing enormous data seamlessly. This data analysis could assist scientists with better knowledge of understanding genetic differences and provide an opportunity to improve these genetic traits for the betterment of human beings.

Methods

Sample collection and DNA extraction

Blood samples of Teddy goat breed (12 individuals) from rural areas of Punjab/Pakistan were collected to extract the genomic DNA through standard protocol and visualized on agarose gel (Figure 1).

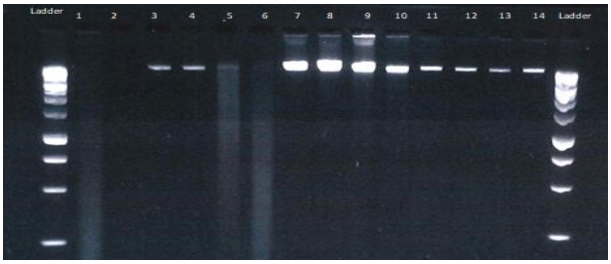

Figure 1: Gel visualization of genomic DNA of 14 Teddy goats. Best 12 were used for pooled sequencing. Both sides of the gel showing 1kb ladder.

Figure 2: True representatives of both breeds Teddy (right) San Clemente (left).

Sequencing runs

Twelve Teddy goat DNA were pooled for Illumine sequencing (HiSeq3000 150bp paired-end/ ~300mio reads). Sequence data was obtained from ENA Project ID PRJEB23815. Reads were quality-filtered using fastq-mcf, mapped with BWA-MEM against ARS1 (San Clemente) reference genome. Further, SAM file converted into bam and sorted with SAM-tools, while duplicated reads were marked using Picard-tools MarkDuplicates feature.

Galaxy platform

Galaxy https://usegalaxy.org/ is web-based open source platform meant for analyzing data, making workflows, its visualizing and sharing data, thus allowing scientists to do computation on their datasets anywhere anytime [7]. All features of galaxy platform are freely available by registering and logging-in. Galaxy presents all-in-one platform aiding researchers to do data analysis spontaneously and in a simplified way with hundreds of bioinformatics tools. In order to get raw data, galaxy could be connected to various genome browsers and databases **(**Figure 3), which shows interface of galaxy with tools list in its left panel, history in right panel and central panel displaying home page.

Figure 3: Graphical user interface (GUI) of web-based Galaxy platform.

Galaxy Tools

FastQC present quality control measures on the raw sequence data, which is essential for further downstream analyses [8]. The FastQC tool can be found under NGS: QC and manipulation tool section on Galaxy Graphical User Interface (GUI). By using this quality control tool, whole genome of Teddy goat can be visualized as graphs (Figure 5). Similarly, FreeBayes is another Galaxy tool used in the current study which follows Bayes' theorem for fabricating vcf files. This probabilistic approach was carried out for each possible genotype in

the pooled samples for observing frequency at each aligned nucleotide, which can ultimately be brought together by measuring differences between the most likely and the second most likely genotypes/observation frequency as a measure of confidence with the highest probability [9]. This tool comes under NGS: Variant Analysis in the Galaxy GUI.

SnpSift toolkit allows filtering and manipulating vcf files. It identifies candidate phenotype-relevant variants and envisages its effect on genes [10]. Filtering variants using arbitrary expressions and annotation was executed using this tool on the obtained vcf file generated by FreeBayes tool. Complete workflow of different Galaxy tools with input and output files used in this study on pooled-Seq data are shown in (Figure 4).

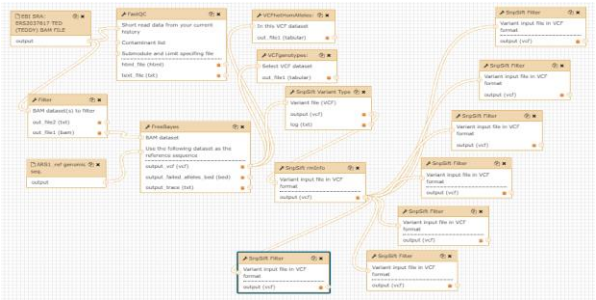

Figure 4: Workflow of Teddy goat variant calling analysis constructed on Galaxy. Tools are mentioned in boxes along with input and output files which are connected with wires.

Databases for sequence retrieval

European Nucleotide Archive (ENA)

ENA https://www.ebi.ac.uk/ena/data/search database archive experimental NGS nucleotide sequence data [11]. Raw data of Teddy goat was retrieved from ENA ID ERS2037817.

National Center for Biotechnology Information (NCBI)

NCBI https://www.ncbi.nlm.nih.gov/ housed total of 45 smaller databases [12], San Clement goat reference genome was fetched from NCBI

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/704/415/GCA_001704415.1_ARS1 (Tool - get data - subtool - uploading file from computer). By clicking paste/fetch data, the link was pasted in the box, convert spaces to tabs in options. After clicking start, reference genome was uploaded in the history panel of the Galaxy GUI.

Ensemble Genome Browser

Ensemble is an integrated database and software system https://asia.ensembl.org/index.html that maintains automatic annotations for genomes. In this study, Variant Effect Predictor (VEP) https://asia.ensembl.org/Tools/VEP a built-in tool set for analysis and annotation of genomic variants was used [13]. It is an upgraded, web-based open source tool that annotates variants in a wide range of designs with simplicity and precision.

Introduction to file formats

Sequencing data received from sequencing platforms in the form of Fastq (Illumina) or Fast5 (ONT) formats, which generally comprises of four lines having sequence reads as well as quality scores of the sequence [14]. Fast QC and Mapping tool are applied on Fastq file which generated "Binary Alignment Map" files. Each row of this file format describes a single alignment of a raw read against the reference genome of ARS1, which is machine readable and has comprehensive details of genome sequencing [15]. BAM file of Teddy goat breed created VCF file using "FreeBayes" a sub tool of NGS: Variant analysis in this study. It starts with the "Header section" consisting of field name, sample names in the data section.

Results

Reads quality checks

The fastq file generated after completing the sequencing run was checked for sequence quality reports using FastQC (Galaxy Version 0.71). Which generated basic text and HTML report that engendered 10 output graphs along with its statistics summary. Statistical basics of input file apprise the total number of sequences processed was 635,357,043 reads. The range of the sequenced length was 30 – 151 kb with GC content of all bases were 42%.

Figure 5 shows summary graphs of per base sequence quality which generates whisker plot. The upper and lower whiskers lies in 10% and 90%, respectively. Per GC sequence content, a graph between numbers of reads vs. GC%/read displaying theoretical distribution [16] proving that no other species has contaminated the sample, and per sequence quality score representing good quality result as the most frequently observed mean quality is above 27, along with distribution of average read quality which is properly tighten in the upper range of the plot.

Functional annotation of variants

Evaluating variants from bam file is the key to explicate the notable characters of Teddy goat breed and to associate these variants with its particular breed characteristics. BAM file were filtered on the basis of position [15] and subjected to vcf file formulation using FreeBayes genetic variant detector, then annotated with putative functional variants using SnpSift tool [17]. Parametric conditions for input of FreeBayes and SnpSift Variant type are displayed in (Figure 6).

Single nucleotide polymorphisms

Annotation revealed 1134 SNPs in which 907 are homozygous and remaining 227 are heterozygous. Consequently, all SNPs include 94% intergenic variants, upstream and downstream gene variant are 3% each, missense and synonymous variant are 0% each and within coding sequences, the ratio of missense variants was 67% and synonymous variants were 33%. Count of overlapped genes and transcripts was 11% and 17% respectively. SNPs on protein coding genes are given in (Table 1).

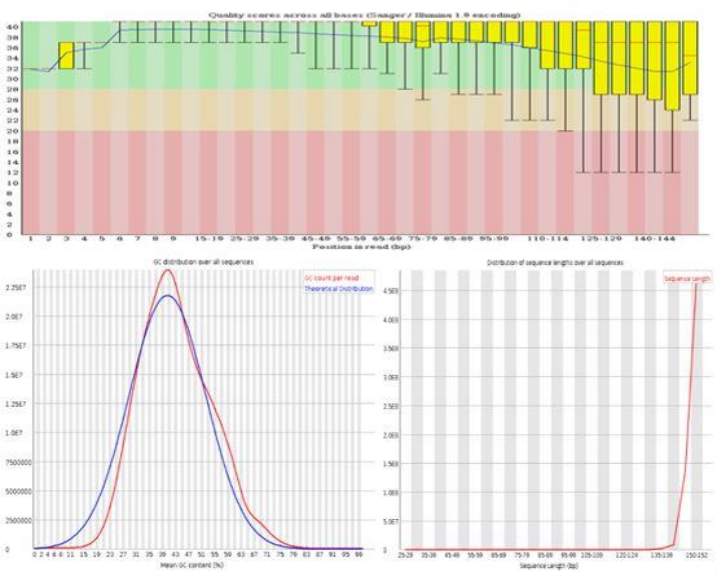

Figure 5: Graphical summary of per base sequence quality and phred score, yellow box of whisker plots displaying inter quartile range of 25-27% (on top). GC count per read are shown as red while theoretical distribution of GC content shown in blue (lower left), and per sequence quality score shown in red (lower right).

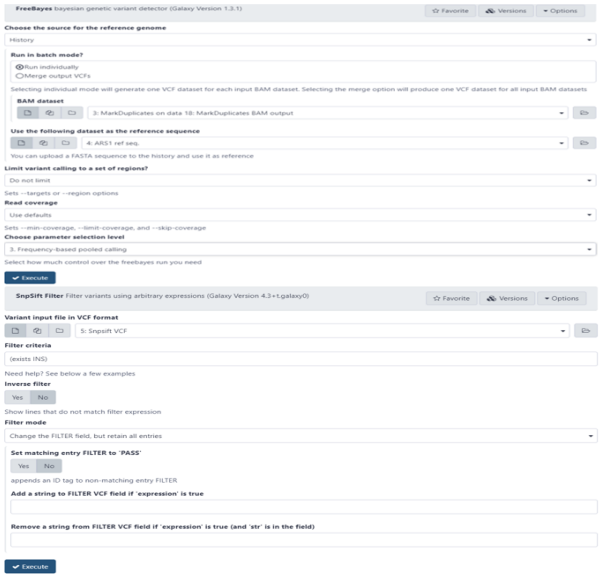

Figure 6: Parametric settings for execution of variant file (top) and variant annotation (lower).

Single Nucleotide Polymorphisms					
Chromosome	**Locus**	**Alt**	**Gene**	**Dist. to transcript**	**Consequences**
LWLT01000035	11; 20; 23; 29; 49; 63; 76; 78; 94; 106; 109	C; A; G; C; C; C; C; C; C; C; G	P2RX6	650; 641; 638; 632; 612; 598; 585; 583; 567; 555; 552	Upstream variant
LWLT01000122	28;30;36;38;52;67;74;100; 104	T; G; G; A; G; G; G; A; G	MICALL2	392; 390; 384; 382; 368; 353; 346; 320; 316	Downstream variant
LWLT01001116	50;90;143	G; C; C	DCX	614;574;521	Downstream variant
LWLT01001501	38;69;84;93;102;132	T; A; C; G; C; G	-	3216,3755;3185,3724; 3170,3709; 3161,3700; 3152,3691; 3122,3661	Upstream variant
LWLT01002501	18;52;60;75;84;128	C; T; C; A; G; C	B4GALNT4	126; 92; 84; 69; 60; 16	Upstream variant
LWLT01002840	9;31;101;107	A; G; A; G	-	178; 156; 86; 80	Downstream variant
LWLT01003190	14;115	G; C	-	4409; 4308	Downstream variant
LWLT01003948	10	T	-	1882	Upstream variant
LWLT01004092	19	C	ARHGAP36	468,468,1587,3067	Downstream variant
LWLT010129626	16;90;93;122	T; G; G; C	-	47; -; -; -	Downstream; Synonymous; Missense
LWLT01013297	7;31;65;75;88;129	G; C; C; T; A; G	-	4962; 4398; 4904; 4894; 4881; 4840	Downstream variant

Table 1: Details of observed SNPs in the Teddy genome in its protein coding genes as compare to ARS1 goat reference genome.

Deletions					
Chromosome	**Locus**	**Alt**	**Gene**	**Dist. to transcript**	**Consequences**
LWLT01002949	**3-5**	**A**	DNAJB8	3080	Downstream variant
Insertions					
LWLT01000302	**20-21**	**TC**	-	2001	Upstream variant
LWLT01002840	**18-20; 28-29; 48-50**	**GGG; TG; CCT**	-	167; 158; 137	Downstream variant
LWLT01003948	**20-22**	**CCC**	-	1870	Upstream variant
Mixed Variants					
LWLT01004092	**48-50**	**CCCC,ACCC**	ARHGAP36	437, 1556, 3036	Downstream gene variant
Multiple Nucleotide Polymorphisms					
LWLT01000122	**5-6**	**AC**	MICALL2	414	Downstream gene variant
LWLT01004092	**39-40**	**CC**	ARHGAP36	447,1566,3046	Downstream gene variant

Table 2: Characteristics count of variants other than SNPs particularly in protein coding genes of Teddy goats.

Other multiple variants

Effects of mixed variants, MNPs and indels of Teddy goat were determined using Variant Effect Predictor. Among these 23 mixed variants (17 Homozygous & 6 Heterozygous) appeared in 01 overlapping gene and in 04 overlapping transcripts, comprising 81% intergenic variant and 19% downstream gene variants. Analysis on MNPs file resulted in 76 variants (67 Homozygous & 9 Heterozygous) with 94% intergenic variant while 6% were downstream gene variants occurring in 02 overlapping genes and 06 overlapping transcripts. Further, 183 insertions appeared in 03 overlapping genes and 04 overlapping transcripts, containing 171 homozygous and 12 heterozygous variants with 96% concurring in intergenic region while 2% of insertions were in downstream and upstream gene region each. Minute proportion of 132 deletions among which (121 homozygous & 12 heterozygous) were present in 01 overlapping genes and 02 overlapping transcripts which lies in 1% downstream region and 99% in intergenic region. Details of variations in protein coding region are given in (Table 2).

Discussion

Generation of high throughput genomic data from NGS technologies has led to the development of bioinformatics software containing various tools for quality check, mapping, variant calling and annotating rare and *de novo* variants and quantifying expression levels of identified genes. Galaxy is a big platform facilitating large genomic dataset analysis and visualization. In this manuscript, different tools are introduced to run for variant calling pipeline analysis on whole genome pooled-seq on Pakistani Teddy goat breeds which were mapped against San Clemente ARS1 reference genome to find out Teddy breed specific variants. This pipeline enabled us to deeply analyze variants and annotating them to have an insight of underlying genes that might be associated with Teddy breed characteristics. Genes were identified by variant effect predictor tool from Ensemble genome browser database. *MICALL2* gene is related to muscle growth [18] while *P2RX6* gene which was found at LWLT01000035 locus in this study showed association to skeletal and muscle development function in Pigs [19]. Furthermore, Galaxy is one of the best platform for analyzing whole genome DNA, RNA and many other renowned pipeline with ready steady bioinformatics tools for smaller data with limited resource scenario. Otherwise, in-house or remote high performance computer (HPC) cluster clouds may also be used with Linux operating systems for the genomics, transcriptomics, proteomics and metabolomics studies. Some of the renowned paid HPCs cloud computing services may be taken from Rescale, Penguin on Demand (POD), AWS, Azure, Alibaba and Google clouds. Setback with these server is unavailability of comprehensive software packages related to bioinformatics analyses, but Galaxy provide the state-of-art availability of variety of latest bioinformatics software and tools to web users for discoveries from their raw multi-omics data without having expertise to work on command line interfaces.

Improvements in bioinformatics and computing power has led to the development of systems and platforms having different types of workflow and pipeline for NGS data analysis, which has enormously facilitated researchers community in analyzing and improving their knowledge of genotype-phenotype relationships. This variant calling pipeline will benefit in understanding the usage of freely available NGS data analyses platforms in limited resources scenario without accessing the expensive in-house/remote cluster and cloud computing facilities. In the mean time, current study provides the opportunity to find out the role of novel whole genome variants in Teddy goat describing different phenotypes rather than by using candidate gene approaches.

Competing interest

All the authors declare that they have no competing interest that can affect the current study.

Authors' Contribution

All authors contributed equally to design, perform, draft and revise this study/manuscript.

References

1. Cock PJ, Grüning BA, Paszkiewicz K, Pritchard L. Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology. PeerJ, (2013); 1e167.
2. Lohmann K, Klein C. Next generation sequencing and the future of genetic diagnosis. Neurotherapeutics, (2014); 11(4): 699-707.
3. Harris D. The distribution and ancestry of the domestic goat; 1962. Wiley Online Library. pp. 79-91.
4. Tahir M, Younas M, Raza S, Lateef M, Iqbal A, *et al*. A study on estimation of heritability of birth weight and weaning weight of Teddy goats kept under Pakistani conditions. Asian-Australasian Journal of Animal Sciences, (1995); 8(6): 595-597.
5. Afzal M, Naqvi A. Livestock resources of Pakistan: present status and future trends. Quart Sci Vis, (2004); 9(1-2): 15-27.
6. Witkowski VM. Study of an endangered species enhancement program in coastal wetlands: public perceptions and management strategies. (1993).
7. Taylor J, Schenck I, Blankenberg D, Nekrutenko A. Using galaxy to perform large-scale interactive data analyses. Current protocols in bioinformatics, (2007); 19(1): 10.15. 11-10.15. 25.
8. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics, (2017); 33(19): 3137-3139.
9. Tange O. Gnu parallel-the command-line power tool. The USENIX Magazine, (2011); 36(1): 42-47.
10. Ruden DM, Cingolani P, Patel VM, Coon M, Nguyen T, *et al*. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in genetics, (2012); 335.
11. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, *et al*. The European nucleotide archive. Nucleic acids research, (2010); 39(suppl_1): D28-D31.
12. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, *et al*. Database resources of the national center for biotechnology information. Nucleic acids research, (2007); 36(suppl_1): D13-D21.
13. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, *et al*. The ensembl variant effect predictor. Genome biology, (2016); 17(1): 122.
14. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research, (2010); 38(6): 1767-1771.

15. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics, (2011); 27(12): 1691-1692.
16. Hastreiter M, Jeske T, Hoser J, Kluge M, Ahomaa K, *et al*. KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis. Bioinformatics, (2017); 33(10): 1565-1567.
17. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, *et al*. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nature methods, (2015); 12(10): 966.
18. Silva VHd (2015) Identification of CNVs in the Nelore genome and its association with meat tenderness: Universidade de São Paulo.
19. Liu X, Du Y, Trakooljul N, Brand B, Muráni E, *et al*. Muscle transcriptional profile based on muscle Fiber, mitochondrial respiratory activity, and metabolic enzymes. (2015); 11(12): 1348.