RESEARCH ARTICLE

Stepwise Detection of Copy Number Variations in Whole Genome Sequence of Buffalo

Aishwarya Dash[1*], Jayakumar Sivalingam[2], Kangabam Bidyalaxmi[1], Kousalya Devi M[1], Nidhi Sukhija[1], Ravi Kumar D[1], Saket Kumar Niranjan[3], Madhu Sudan Tantia[3], Ishwar Dayal Gupta[1]

Abstract

Genetic variants cause changes in genetic composition and are accountable for uniqueness of genomes of individuals. Copy number variations (CNVs) cause genetic variations due to loss or gain of DNA segments varying from 50 bp to several mega base pairs (Mb) compared with a reference genome. These unbalanced structural variants are heritable and have potentially greater effect than SNPs. The present study was conducted on whole genome resequencing (WGS) data of swamp buffalo for identification of CNVs. Sample of swamp buffalo was collected from Manipur and karyotyped and confirmed to be of swamp type. After quality and quantity checking it was sent for sequencing with Illumina Hiseq 2000 platform. Paired end reads of WGS data were analysed using Read depth (RD) based approach aligning to the reference Mediterranean riverine buffalo assembly. CNVs were identified using CNVnator with standardized parameters and total 587 CNVs were found in swamp buffalo. Among CNV events, 405 deletions and 182 duplications were identified. These variations comprised of 5.82% of reference genome. This was the first Genome-wide CNV study in indigenous swamp buffalo with reference to buffalo genome using Read Depth based approach.

Key words: Buffalo, Copy number variations (CNVs), Read Depth, Whole genome sequencing (WGS)

Ind J Vet Sci and Biotech (2023): 10.48165/ijvsbt.19.3.07

Introduction

Copy number variations (CNVs) are unbalanced structural variations due to loss or gain involving DNA segments varying from 50 bp to several mega base pairs (Mb) compared with a reference genome. It is a kind of genetic variants affecting larger segments, which is resulted due to loss by deletion or gain by duplication, insertional translocation (Mills *et al*., 2011). As CNVs involve a large segment of DNA, they have a potentially greater effect than SNPs though SNPs are common polymorphism. They are thought to be more heritable sequence variations between individuals considering total number of nucleotides (Ghosh *et al*., 2014). There is significant progress in studies on CNVs of various livestock genomes of cattle (Zhou *et al*., 2022), buffalo (Strillacci *et al*., 2021), sheep (Salehian-Dehkordi *et al*., 2021), goat (Nandolo *et al*., 2021), pig (Qiu *et al*., 2021) and chicken (Fernandes *et al*., 2021). CNV identification in buffalo is merely studied and existing studies for CNV detection in buffaloes are performed with reference to cattle genome. Several methods have been used to identify CNVs like array comparative genomic hybridisation (aCGH), SNP arrays and Next generation sequencing (NGS) (Pinto *et al*., 2011; Gao *et al*., 2017). NGS data can be used to explore structural variations with four basic strategies such as, Paired-end mapping (PEM) or Read Pair (RP), Split Read (SR), Read Depth (RD), Sequence Assembly or de novo Assembly of Genome (AS) (Pirooznia *et al*., 2015; Gao *et al*., 2017). Among different NGS techniques, Read Depth (RD) based assessment of CNVs is based on the hypothesis that there is correlation between depth of coverage and copy number of one genomic region and helps in finding large size CNVs and exact copy number of events. Therefore, the present investigation was carried out using the whole genome sequence (WGS) of swamp buffalo genome with respect to Mediterranean riverine buffalo genome.

[1]Division of Animal Genetics and Breeding, ICAR-National Dairy Research Institute, Karnal -132001, Haryana, India.

[2]ICAR-Directorate of Poultry Research, Hyderabad, Telangana, India.

[3]ICAR-National Bureau of Animal Genetics and Breeding, Karnal-132001, Haryana, India

Corresponding Author: Aishwarya Dash, Division of Animal Genetics and Breeding, ICAR-National Dairy Research Institute, Karnal -132001, Haryana, India, e-mail: dashaishwarya@gmail.com

How to cite this article: Dash, A., Sivalingam, J., Bidyalaxmi, K., Kousalya Devi, M., Sukhija, N., Ravi Kumar, D., Niranjan, S.K., Tantia, M.S., & Gupta, I.D. (2023). Stepwise Detection of Copy Number Variations in Whole Genome Sequence of Buffalo. Ind J Vet Sci and Biotech. 19(3), 30-33.

Source of support: Nil

Conflict of interest: The author(s) declare that there is no conflict of interest.

Submitted: 13/02/2023 **Accepted:** 01/04/2023 **Published:** 10/05/2023

Materials and Methods

Whole Genome Re-sequencing Data

Blood samples of 20 swamp buffaloes including 6 males and 14 females were collected from the field at Senapati, Chandel and from the Swamp Buffalo Breeding Farm at

© The Author(s). 2023 Open Access This work is licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International License.

Wabagai village of Thoubal district of Manipur. The random blood sampling of animals was performed in accordance with the relevant guidelines and regulations as approved by Institutional Animal Ethics Committee (IAEC) of National Bureau of Animal Genetics Resources (NBAGR), Karnal. Samples were karyotyped using blood lymphocyte culture technique and it was confirmed that all samples were of swamp type. Genomic DNA of each sample was isolated and quality and quantity were measured by 0.8% agarose gel electrophoresis and NanoDrop technique, respectively. One male sample collected from the field of Senapati district was sent for whole genome resequencing of 2X150 bp paired-ends using Illumina Hiseq 2000 platform. After obtaining WGS data of swamp buffalo, standardized workflow (Fig. 1) was performed as follows.

Quality check of reads: Paired end reads of FASTQ files received from high throughput Illumina sequencing pipelines, were analysed using FastQC version 0.11.8 (Andrews, 2010) for assessing quality.

Adapter removal: Quality control (QC) was performed with Trimmomatic-0.39 (Bolger *et al*., 2014) to remove adapters and to trim reads with Phred-score less than 20 at the beginning (LEADING:20) and end of the sequence (TRAILING:20) and reads shorter than 50 bp.

Alignment: Processed reads were aligned with Mediterranean riverine buffalo reference genome assembly (UOA_WB_1) (PRJNA471901) using BWA-MEM version 0.7.17-r1188 (Li and Durbin, 2009) with the default parameters. The output SAM (Sequence Alignment Map) file then was converted to BAM (Binary Alignment Map) file and then sorting was performed using SAMtools version 1.10 (Li *et al*., 2009).

PCR duplicate removal and RD estimation: PCR duplicates are obtained on sequencing two or more copies of the exact same DNA fragment in sequence reads due to PCR amplification bias. They could affect the CNV analysis, so these were marked and removed using SAMtools Markdup, and then depth of coverage was calculated using SAMtools depth (Li *et al*., 2009).

CNV detection: CNVnator version 0.4.1 was run on the BAM file to identify CNVs (Abyzov *et al*., 2011) with window size (bin size) of 1 Kb and other default parameters. Steps followed to perform CNV calling using CNVnator software is outlined in Fig. 2. To set bin size, the ratio of average RD signal to its standard deviation with 4-5 was detected using CNVnator-eval on our sample with different bins (Coghlan, 2013).

Filtration of CNVs: Quality control was performed on the raw CNVs with filtering criteria p-value (e-val1) < 0.01, q_0< 0.5. The p-value < 0.01 indicated region between two calls doesn't have the same CNV and q_0 showed a fraction of mapped

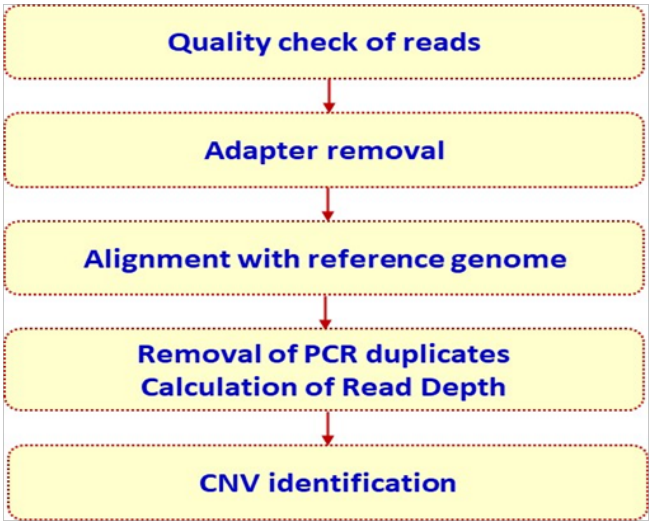

Fig. 1: Workflow for identification of CNVs

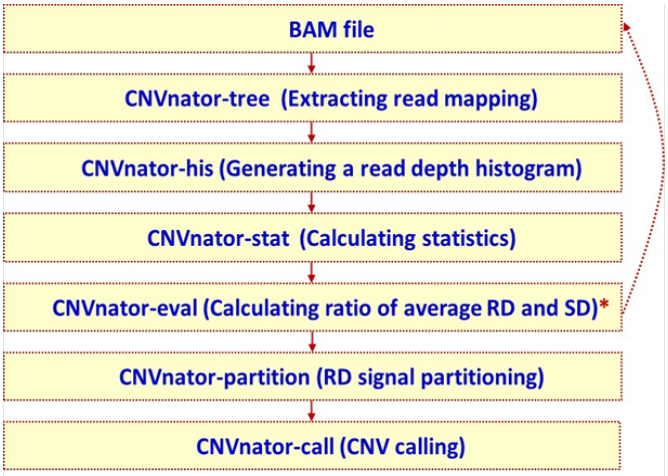

Fig. 2: Flowchart for CNVnator software
(*Choosing bin size to keep ratio of average Read Depth (RD) and its Standard Deviation (SD) is around 4)

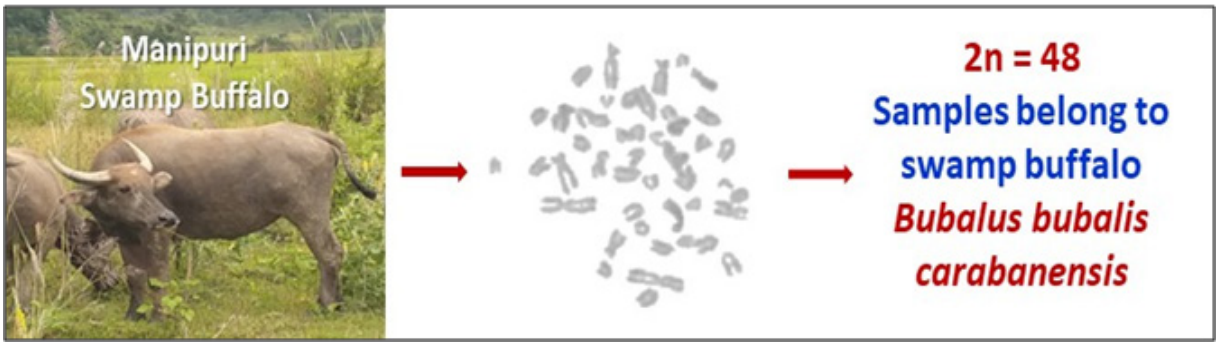

Fig. 3: Karyotyping of swamp buffalo

reads with zero quality. CNVs overlapped with unplaced chromosomes (chrUN) were separated (Gao *et al*., 2017).

Results and Discussion

Sequencing and Mapping

Samples karyotyped using standard lymphocyte culture techniques were found to be of swamp type (Fig. 3). Sequence of Illumina Hiseq paired end reads (151 bp) was obtained and quality control was performed. Low quality bases and adapter were trimmed and put for alignment with UOA_WB_1 riverine genome assembly. PCR duplicates were marked and removed from the bam file. Out of total reads, 98.19% were mapped to reference genome. Depth of coverage of the sample was 7.92X, which was sufficient for CNV identification in terms of read depth as minimum 4X coverage data was required (Sudmant *et al*., 2010; Bickhart *et al*., 2012; Gao *et al*., 2017).

Identification of CNVs

In swamp buffalo, total raw 1331 CNVs were detected. Out of which 856 CNVs were present in chromosomes and 475 CNVs were overlapped with scaffolds. After the quality control of CNVs present in chromosomes, a total of 587 CNVs were obtained (Table 1). Filtered CNVs occupied 157.1 Mb, which covers 5.82% of the reference genome. Duplication events were identified in 182 CNVs with a total size of 144.59 Mb, whereas the deletion events in 405 CNVs of 12.51 Mb size in the genome of swamp buffalo (Table 2). Deletion CNV events were more common than duplication events (Table 3).

Table 1: Summary of CNV events in buffalo

CNV events	Counts
Total CNVs	1331
Unplaced Chr CNVs	475
CNVs (Unfiltered)	856
CNVs (Filtered)	587
Deletion	405
Duplication	182

Table 2: Size and coverage of copy number variations of swamp buffalo

Events	Size	Coverage
CNVs	157.1 Mb	5.82%
Deletion	144.59 Mb	5.36%
Duplication	12.51 Mb	0.46%

Size of Reference Genome (Riverine Mediterranean buffalo = 2.7 Gb)

Zhang *et al*. (2014) identified 163 CNVRs from 3 buffalo samples using cross species CGH Array approach, which comprised 1.44% of cattle reference genome. In 14 buffaloes, 5191 and 8548 CNVs were identified with an average of 371 and 610 CNVs by JaRMS calls and cn.mops software based on RD based approach (Li *et al*., 2019). A total of 1,344 CNV regions were detected across 15 riverine buffaloes amounting to 59.8 Mb size or 2.2% of the cattle genome using mrFAST/mrsFAST based on RD method (Liu *et al*., 2019). Li *et al*. (2019) observed more deletion events, whereas Liu *et al*. (2019) identified more gain CNV events or duplications.

Table 3: Chromosome-wise CNV events in swamp buffalo

Chromosome	Deletion	Duplication	CNVs
1	8	5	13
2	22	5	27
3	16	12	28
4	19	16	35
5	19	8	27
6	15	5	20
7	11	4	15
8	10	6	16
9	12	15	27
10	13	3	16
11	5	8	13
12	5	1	6
13	24	16	40
14	2	6	8
15	8	3	11
16	24	14	38
17	4	5	9
18	18	16	34
19	5	2	7
20	11	2	13
21	4	0	4
22	2	0	2
23	6	1	7
24	1	3	4
X	141	26	167
Total	405	182	587

Distribution of CNVs

The CNVs ranged from 5 kb to 5836 kb. The average sizes of deletion and duplication events in the CNVs were 400.25 kb and 55.94 kb, respectively. The median of CNV events in swamp buffalo for deletion was 45 kb and for duplication was 35 kb.

Conclusions

This was the first genome-wide copy number variation (CNV) study attempted in Indian swamp buffalo by whole genome resequencing method. Copy number variations were detected using Read Depth based approach after aligning with Mediterranean riverine buffalo genome assembly. Total 587 CNVs were identified using CNVnator software with 405

and 182 CNVs involved in deletion and duplication events. These structural variations were found to be of 5 kb in size or as long as 5 Mb. X chromosome was seen to gather maximum of the CNV events. The detected CNVs amounted to 157.1 Mb length in swamp buffalo that accounted for 5.82% of the reference buffalo genome.

Acknowledgement

Authors express their thankfulness to the Directors of ICAR-NBAGR and ICAR-NDRI for the support to carry out this research work. Financial assistance by ICAR-NBAGR to generate the data under the institute funded project "Reference based Genome assembly of Indian swamp buffalo" is thankfully acknowledged. Financial assistance provided by ICAR-NDRI and computational assistance at ICAR-SBI regional centre are thankfully acknowledged.

References

Abyzov, A., Urban, A.E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research, 21*(6), 974-984.

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

Bickhart, D.M., Hou, Y., Schroeder, S.G., Alkan, C., Cardone, M.F., Matukumalli, L.K., *et al.* (2012). Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Research, 22*(4), 778-790.

Bolger, A.M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120.

Coghlan, A. (2013). Avrilomics: Using CNVnator to find copy number variation. *Avrilomics*.

Fernandes, A.C., da Silva, V.H., Goes, C.P., Moreira, G.C.M., Godoy, T.F., Ibelli, A.M.G., *et al.* (2021). Genome-wide detection of CNVs and their association with performance traits in broilers. *BMC Genomics, 22*(1), 1-18.

Gao, Y., Jiang, J., Yang, S., Hou, Y., Liu, G.E., Zhang, S., *et al.* (2017). CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics, 18*, 1-12.

Ghosh, S., Qu, Z., Das, P.J., Fang, E., Juras, R., Cothran, E.G. *et al.* (2014). Copy number variation in the horse genome. *PLoS Genetics, 10*(10), e1004712.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25(14): 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al.* (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079.

Li, W., Bickhart, D.M., Ramunno, L., Iamartino, D., Williams, J.L., & Liu, G.E. (2019). Comparative sequence alignment reveals river buffalo genomic structural differences compared with cattle. *Genomics, 111*(3), 418-425.

Liu, S., Kang, X., Catacchio, C.R., Liu, M., Fang, L., Schroeder, S.G., *et al.* (2019). Computational detection and experimental validation of segmental duplications and associated copy number variations in water buffalo (*Bubalus bubalis*). *Functional and Integrative Genomics, 19*(3), 409-419.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., *et al.* (2011). Mapping copy number variation by population-scale genome sequencing. *Nature, 470*(7332), 59-65.

Nandolo, W., Mészáros, G., Wurzinger, M., Banda, L.J., Gondwe, T.N., Mulindwa, H.A., *et al.* (2021). Detection of copy number variants in African goats using whole genome sequence data. *BMC Genomics, 22*(1), 1-15.

Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., *et al.* (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nature Biotechnology, 29*(6), 512-520.

Pirooznia, M., Goes, F.S., & Zandi, P.P. (2015). Whole-genome CNV analysis: Advances in computational approaches. *Frontiers in Genetics, 6*, 138.

Qiu, Y., Ding, R., Zhuang, Z., Wu, J., Yang, M., Zhou, S., *et al.* (2021). Genome-wide detection of CNV regions and their potential association with growth and fatness traits in Duroc pigs. *BMC Genomics, 22*(1), 1-16.

Salehian-Dehkordi, H., Xu, Y.X., Xu, S.S., Li, X., Luo, L.Y., Liu, Y.J., *et al.* (2021). Genome-wide detection of copy number variations and their association with distinct phenotypes in the world's sheep. *Frontiers in Genetics, 12*, 670582.

Strillacci, M.G., Moradi-Shahrbabak, H., Davoudi, P., Ghoreishifar, S.M., Mokhber, M., Masroure, A.J., *et al.* (2021). A genome-wide scan of copy number variants in three Iranian indigenous river buffaloes. *BMC Genomics, 22*(1), 1-14.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), 75-81.

Zhang, L., Jia, S., Yang, M., Xu, Y., Li, C., Sun, J., *et al.* (2014). Detection of copy number variations and their effects in Chinese bulls. *BMC Genomics, 15*, 1-9.

Zhou, J., Liu, L., Reynolds, E., Huang, X., Garrick, D., & Shi, Y. (2022). Discovering copy number variation in dual-purpose Xin Jiang Brown cattle. *Frontiers in Genetics, 12*, 747431.