Vietnam Journal of Science and Technology xx (20xx)

doi:10.15625/2525-2518/20054

Comparison of Galaxy and Unix tools for analyzing the exome sequencing data from syndactyly abnormalities

Nguyen Thy Ngoc [*], Huynh Minh Huong

University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay District, Ha Noi, Viet Nam

[*]Email: *nguyen-thy.ngoc@usth.edu.vn*

Received: 12 February 2024; Accepted for publication: 17 September 2024

Abstract. Syndactyly is a congenital limb abnormality, which manifests as the fusion of digits due to incomplete separation during embryonic development, and its pathogenesis involves intricate genetic and molecular processes. Since Exome sequencing has gained widespread utilization as an invaluable tool for exploring genetic disorders during prenatal development, the Bioinformatic platforms, such as GALAXY and UNIX, play a central role in the analysis process of exome sequencing data, facilitating precise identification and interpretation of genetic variations linked to congenital abnormalities. In this study, we conducted a comparative analysis of exome sequencing data from a 1.5-year-old syndactyly patient using two platforms: GALAXY and UNIX. The UNIX platform identified a total of 275,572 variants, and the GALAXY platform identified 140,291 variants when compared with the Grch38/hg38 reference genome. A comparative analysis identified 126,848 common variants between the platforms. After filtration with the 200 syndactyly-related genes, 1,345 variants were remained. The distribution of these 1,345 variants spans the entirety of the patient's genome, with focal concentrations observed on specific chromosomes including chromosomes 2, 4, and 11. Concurrently, within the top 200 genes implicated in syndactyly, the genes *FRAS1*, *CACNA1C*, *GLI2*, and *NOTCH1* exhibit the highest frequency of variants. The significance of bioinformatic methods on the discovery of genetic variants linked to syndactyly was highlighted in this work. Galaxy is suggested for users looking for a more user-friendly interface with a platform that is more accessible and repeatable, especially for smaller datasets or standard studies. Unix, on the other side, is recommended for users who need greater performance, adaptability, and the capacity to manage complicate analysis and massive amounts of data. These data emphasized the impact of the chosen analytical platform on genetic variation detection in congenital limb abnormalities, provided critical insights into the selection of bioinformatic tools for optimizing exome sequencing workflows in the context of limb malformations, contributed to advancements in genetic research and diagnostic methodologies.

Keywords: Galaxy, exome sequencing, high performance computing, syndactyly, pipeline.

Classification numbers: 1.2.5, 1.4.3, 4.8.5.

1. INTRODUCTION

Syndactyly is a congenital anomaly characterized by the fusion of fingers or toes through shared skin, soft tissue, and sometimes even cartilage and bone. This represents a relatively

common congenital malformation, with an incidence in newborns ranging from 1/2000 to 1/3000 births [1], occurring more frequently in males with a ratio of 2 males to 1 female [2, 3]. Syndactyly is primarily caused by improper apoptosis (programmed cell death) during embryonic development, leading to insufficient separation of digits. In many cases, these digital skeletal anomalies are not isolated but are associated with syndromes such as Apert syndrome, which leads to abnormal skull development and impacts the intellectual development of affected individuals [4], Poland syndrome, Möbius syndrome, or Weyer Ulnar Ray syndrome [5]... Despite the clinical phenotypes of this anomaly being well-described in numerous studies, and research indicating a significant role of genetic factors in its manifestation, there is a scarcity of publications specifying the responsible genes for this anomaly. Some genes have been identified in patients with syndactyly, including homeobox D13 (*HOXD13*) [6], gap junction protein alpha 1 (*GJA1*), *GLI3* gene [7], limb development membrane protein 1 (*LMBR1*), lipoprotein receptor-related protein 4 (*LRP4*), and fibroblast growth factor 16 (*FGF16*) [8], However, the genes or gene groups responsible for syndactyly subtypes I-a, I-b, I-d, II-c, VI, VIII-b, and IX remain unclear. Understanding the intricate molecular and cellular processes involved in limb development and how disruptions in these processes contribute to syndactyly is crucial for physicians in advancing our knowledge of the condition, diagnosis, marital counseling, and genetic counseling for affected individuals. Additionally, it forms the foundation for timely intervention measures. The specific pathogenic mechanisms can vary based on the underlying genetic mutations and environmental influences in individual cases.

The next-generation sequencing (NGS) technology has provided significant advancements in the field of genetic testing. Progress in sequencing techniques has generated a vast amount of genetic data, leading to the need for managing and analyzing large raw data sets. Applications of NGS such as whole genome sequencing, or whole exome sequencing have been widely used to explore the genetic interactions and mechanisms of different prenatal disorders since it offers a comprehensive and efficient approach to explore the genetic interactions and underlying mechanisms of various prenatal disorders by selectively sequencing the protein-coding regions of the genome, providing insights

into potential causative variants [9]. However, many bioinformatics analysis tools (pipelines) nowadays require a high-performance computing (HPC) platform and programming expertise on the UNIX platform, posing challenges for researchers inexperienced in HPC to construct and manage pipelines. In this study, we utilize GALAXY as a freely accessible public server with a flexible bioinformatics platform and a user-friendly interface for those with little to no programming experience on UNIX. We employ GALAXY to analyze the entire Exome sequencing dataset of patients with syndactyly. Furthermore, GALAXY facilitates the integration of numerous rich toolsets, enhancing its versatility in handling various sequencing datasets [10].

2. MATERIALS AND METHODS

2.1. Description of patient and data

A male patient, aged 1.5 years, presented with syndactyly of the little toe and the adjacent toe. The specific condition of the patient and family members has been previously detailed in a study [7]. Patient sampling was conducted under the Ethics Committee of the Central Pediatric Hospital's Biomedical Research, in accordance with Decision No. 564/BVNTW-VNCSKTE. The whole exome sequencing was conducted using SureSelectXT Library Prep Kit with 151 bp read length on Illumina platform. 10,683,606,662 bases were

obtained from 70,752,362 paired-end read sequences. Of these, 97.26 % of reads had a quality score (Q score) exceeding 20, with 92.71 % surpassing a Q score of 30. The GC base pair percentage in the exome data was 51.82 %.

2.2. Research methods

2.2.1. The GALAXY pipeline

The acquired raw exome sequencing data underwent analysis utilizing the GALAXY platform (https://usegalaxy.eu/). The removal of adapters and trimming of bases from the terminal regions with the quality score less than 10 was executed using Trimmomatic. Subsequently, the data underwent mapping and alignment procedures through BWA-MEM, utilizing the reference index of the human genome version GRCh38/hg38.

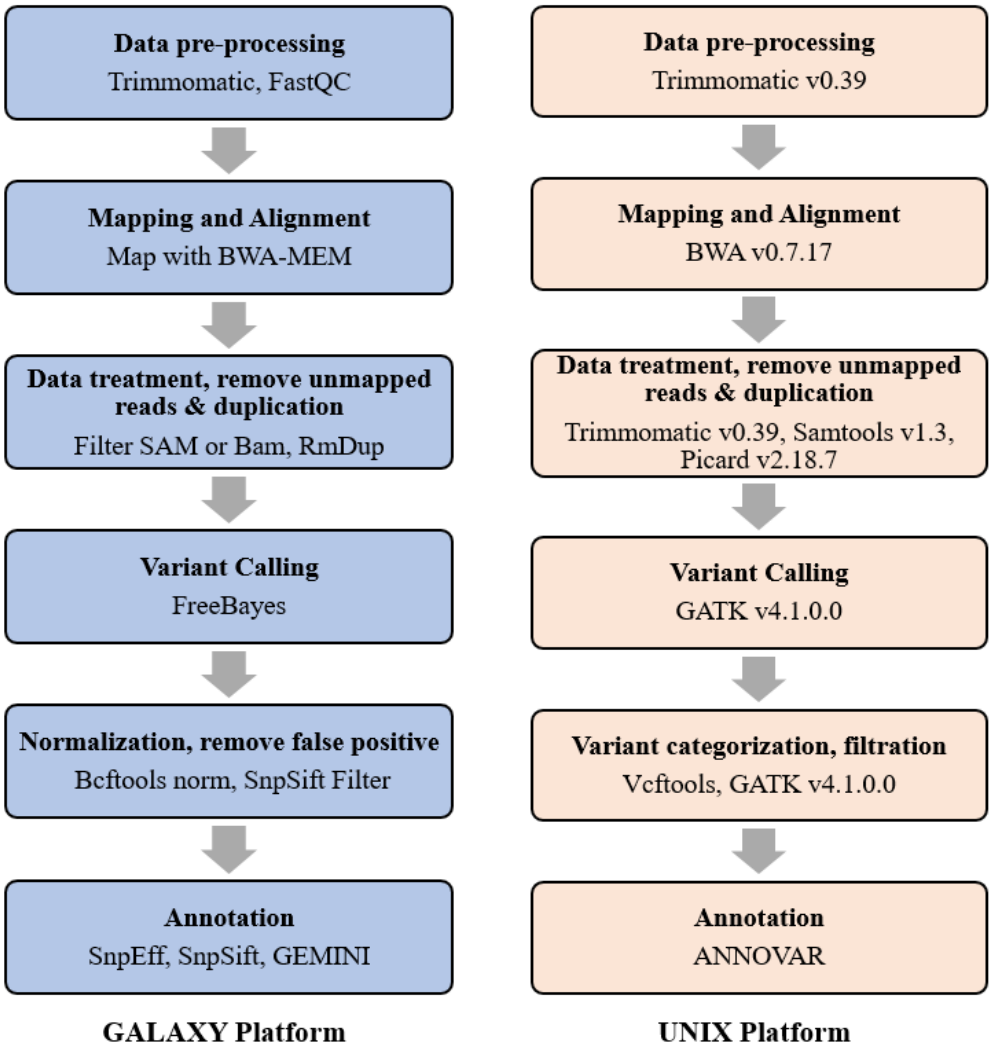

Figure 1. The bioinformatic pipelines for exome sequencing analysis with GALAXY and UNIX. The analysis steps are annotated in bold font, and the software used for each step is indicated in regular font.

The Filter SAM or Bam, output SAM or BAM tool was used to refine and extract specific subsets of data, and filter unmapped reads. Duplicate reads from Next-generation sequencing dataset was removed using RmDup. To call out genetic polymorphisms, the FreeBayes tools was launched using the hg38 reference genome. The Bcftools norm tool was employed to address multiallelic variants by splitting them and to left-align indels within the variant data. Following this normalization step, the mitigation of false-positive variant calls was accomplished through the utilization of SnpSift filter. Specifically, variants with a quality by depth (QD) metric greater than 2 were retained, thereby enhancing the accuracy and reliability of the variant calls by filtering out those with lower quality measurements. The three tools SnpEff, SnpSift and GEMINI were applied to annotate and predict the functional effects of the obtained genetic variants.

2.2.2. The UNIX pipeline

For the UNIX tools, the exome sequences of the obtained reads were removed bad reads and adapters by Trimmomatic v0.39, aligned and mapped to the human reference genome version GRCh38/hg38 using the Burrows-Wheeler Aligner (BWA) software version 0.7.17. Low-quality reads or reads with unaligned positions were filtered out using Trimmomatic software v0.39 and Samtools v1.3. PCR amplification artifacts, such as duplicated reads, were marked and removed using Picard software v2.18.7. After comparing with the reference genome, the polymorphic points and the different mutations were identified using the Genome Analysis Toolkit (GATK) v4.1.0.0 software. The data then was separated into single nucleotide polymorphism (SNP) or insertion/deletion (indel) variants by Vcftools and filtered by GATK to remove bad call. The remaining mutation points were annotated with gene positions using the ANNOVAR software. The bioinformatic pipeline construction was based on the methodology described in the study by Van der Auwera et al [11]. The bioinformatic workflows to analyze exome sequencing data based on GALAXY and UNIX were described in Figure 1.

2.2.3. The list of potential causative genes in syndactyly

The list of top 200 genes with the best syndactyly-related scores were obtained from the Genecards human gene database to filter possible causative gene variations (https://www.genecards.org/) (Annex 1). These genes have been demonstrated to be involved in the pathogenesis of syndactyly in many studies [12, 13].

3. RESULTS AND DISCUSSION

In the analysis of exome sequencing data using the GALAXY platform, a total of 140,291 variants were identified, comprising 129,792 SNPs and 10,499 indels. In the other hand, the UNIX platform yielded a nearly two-fold increase in the number of variants called from the hg38 reference genome: 275,572 variants, with 248,392 SNPs and 27,180 indels. Subsequent filtering based on a curated list of the top 200 genes associated with syndactyly-related scores resulted in the retention of 1,506 variants from the GALAXY platform (consisting of 1,401 SNPs and 105 indels), while the UNIX platform retained 3,481 variants (comprising 3,134 SNPs and 347 indels). The diverse types of detected variants and their distribution across different regions (coding region, intron, intergenic, splicing region) were detailed in Table 1. These results showed that the variants obtained from the UNIX platform were more abundant compared to the GALAXY system, which may be attributed to the rigid quality control

procedures for read segments implemented by the GALAXY platform, with quality control steps at each stage that eliminate reads with poor quality.

Table 1. Number of variants detected with the GALAXY and UNIX platforms.

	GALAXY Platform		**UNIX Platform**	
	SNP	Indel	SNP	Indel
Before filtration with the 200 syndactyly-related genes				
Total variants	129,792	10,499	248,392	27,180
Coding region	9,963	429	20,962	482
Splice region	2,256	442	74	39
Non-coding exon	3,330	159	3,422	322
Novel variants	3,259	288	7,315	2,347
After filtration with the 200 syndactyly-related genes				
Total variants	1,401	105	3,134	347
Coding region	258	2	266	4
Splice region	30	5	1	1
Non-coding exon	5	0	11	0
Novel variants	3	2	34	28

Subsequent comparisons indicated that among 289,015 variants identified by GALAXY or UNIX platforms before filtration with the list 200 syndactyly-related genes, 126.848 variants were common between the two pipelines, including 117,339 SNPs and 9,509 indels. After filtration, the two platforms showed 3,642 variants. Among those 1,345 variants were common, with 1,257 SNPs and 88 indels (Figure 2).

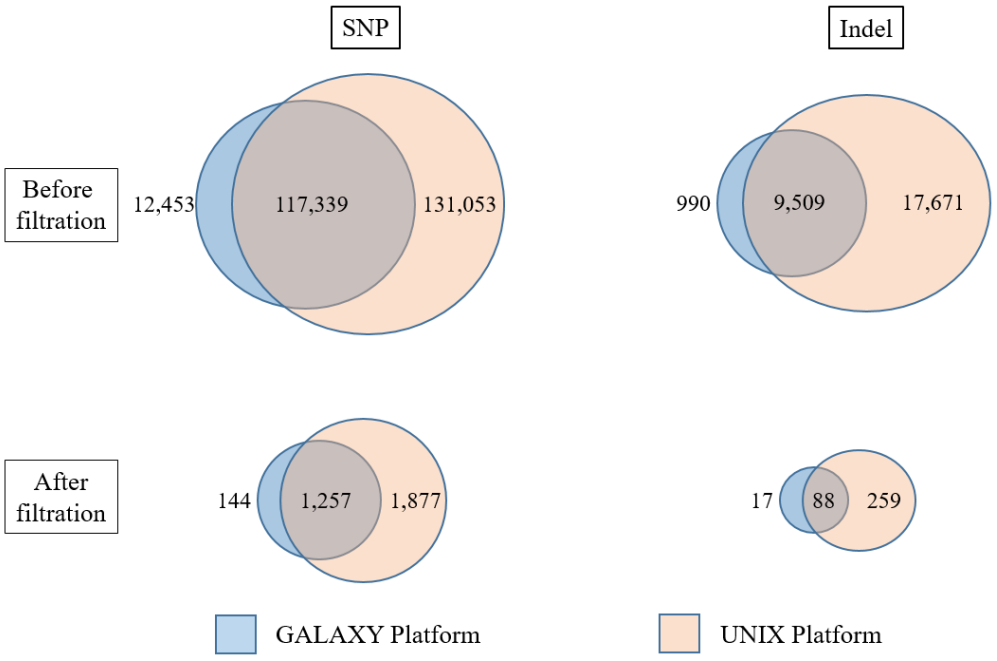

Figure 2. Venn diagrams illustrated the common variants, including SNPs and indels, between the two platforms GALAXY and UNIX before and after filtering with 200 syndactyly-related genes.

Location analysis showed that the majority of variants were situated on chromosomes 2, 4, and 11, comprising 131, 128, and 121 variants, respectively. Conversely, a relatively lower number of variants were observed on chromosomes 8, 21, and X, accounting for 21, 7, and 22 variants (Figure 3). The chromosomal distribution of the 1,345 shared variants between GALAXY and UNIX platforms is illustrated in Figure 3.

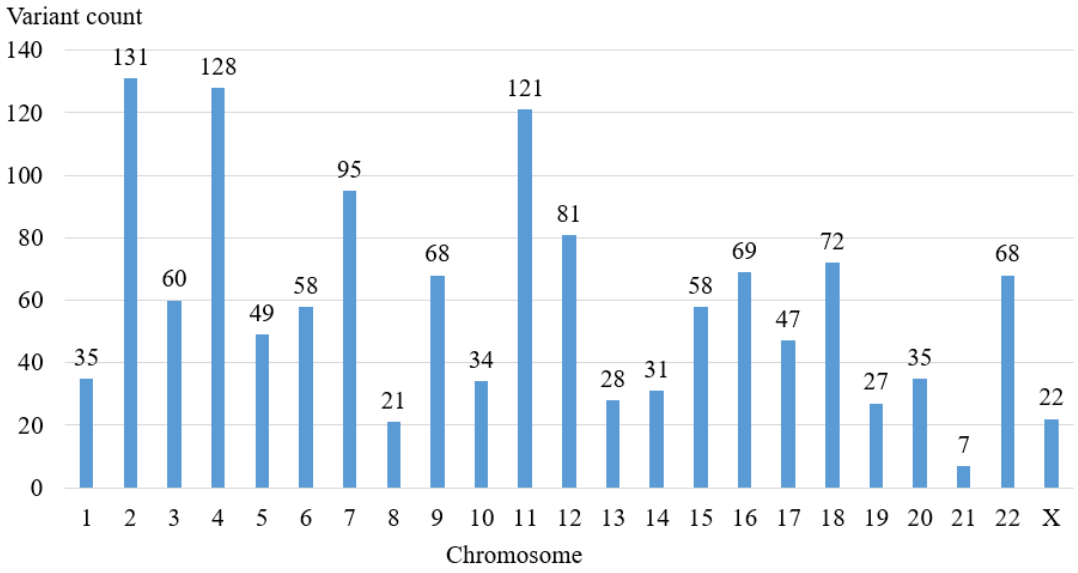

Figure 3. Chromosomal distribution of 1,345 shared variants between GALAXY and UNIX platforms.

Further analysis of these 1,345 shared variants indicated that, within the list of top 200 genes associated with syndactyly, we detected numerous variants located in the *FRAS1* gene (70 variants), *CACNA1C* gene (43 variants), *GLI2* gene (42 variants) and *NOTCH1* (33 variants) (Figure 4A). Among the top 20 genes associated with syndactyly, the highest variant count genes were *CACNA1C* gene (43 variants), *LRP4* gene (27 variants), *GLI3* (11 variants) and *FGFR2* (14 variants), whereas no variant was observed in the *GJA1*, *HOXD13*, *NECTIN4* and *CCNQ* gene (Figure 4B).

The *FRAS1* gene encodes a protein that is essential for the formation of the extracellular matrix, which is a network of proteins and other molecules outside cells that provide structural support. This extracellular matrix is particularly important in skin and limb formation during embryonic development. A nonsense mutation resulting in the loss of *FRAS1* function has been implicated in the disruption of inter-digital apoptosis, leading to syndactyly in a mouse model. [14]. The *CACNA1C* gene, in the other hand, encodes for the alpha-1C subunit of a voltage-gated calcium channel, which plays a key role in regulating calcium ion flow into cells. Disruptions in these pathways can lead to abnormalities in tissue and organ formation. For instance, Timothy syndrome, a rare genetic disorder caused by mutations in *CACNA1C*, includes syndactyly as one of its features [15]. While the two genes *GLI2* and *NOTCH1* are involved in the Hedgehog and Notch signaling pathways, respectively. These pathways play significant roles in cell differentiation, proliferation, apoptosis and the proper development of various tissues, including limbs patterning [16, 17].

Exome sequencing is widely used to find pathogenic variations for the identification of suspected genetic disorders like syndactyly. Nevertheless, errors in data processing, alignment,

and variant calling stages of NGS bioinformatic pipelines may give rise to false positives or inaccurately called variants [18]. Therefore, results from different analysis pipeline tools is necessary to decrease false positive errors and facilitate a comprehensive investigation into the genetic architecture of syndactyly.

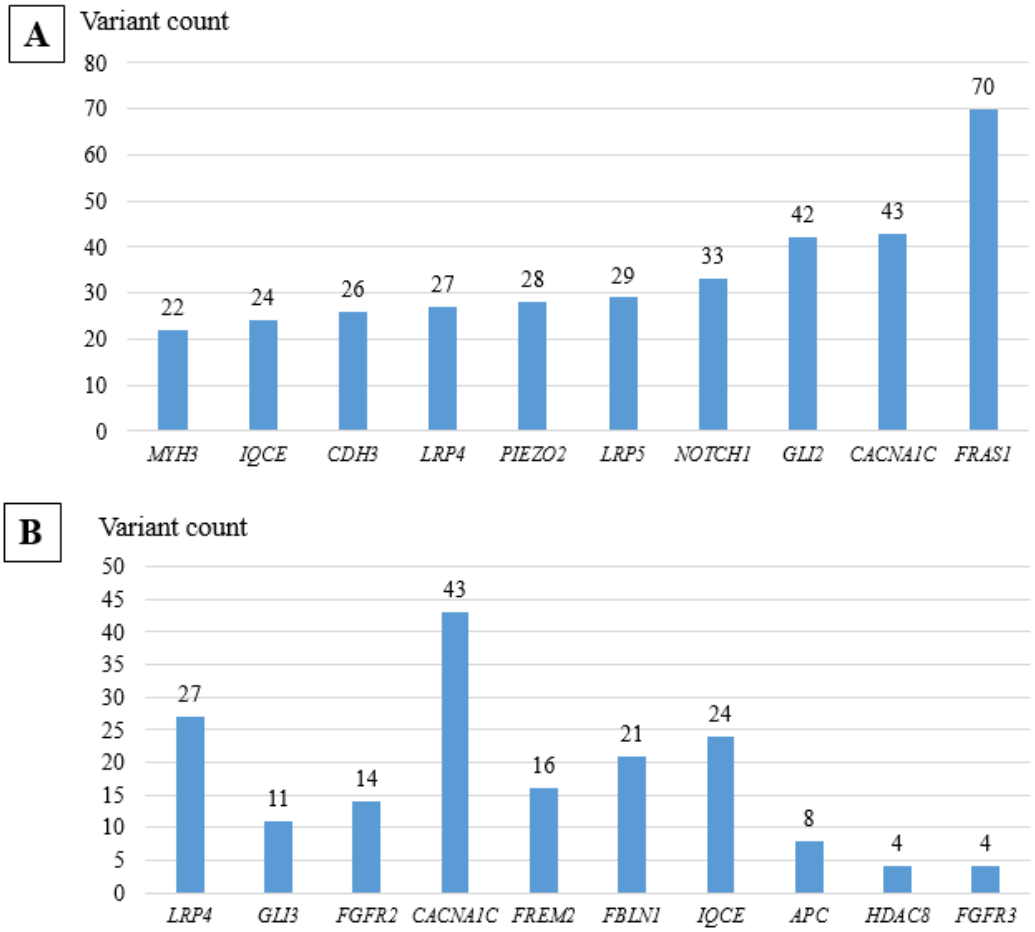

Figure 4. Statistics of genes with the highest variant count among the top 200 genes (A) and top 20 genes associated with syndactyly (B)

Galaxy represents a robust and user-friendly web-based platform designed for scientific data analysis. The execution of analysis steps involves the utilization of Galaxy tools, which provide instructions on the translation of parameters for command-line software into an accessible web interface [19]. The backend architecture of Galaxy is designed to seamlessly interface with diverse cloud and high-performance computing (HPC) environments. This functionality ensures the availability of requisite computational resources for executing computationally intensive analyses. Notably, end users can perform these analyses with minimal requirements through accessibility to a web browser. There are different tools that can be seamlessly utilized both from the Unix command line and within the Galaxy platform for enhanced accessibility and workflow integration such as Bowtie, Samtools, (computational genomics analysis toolkit (CGAT) [20]. However, the UNIX command system has some disadvantages such as: requires manual management of tools, dependencies, and workflow

construction, posing challenges for researchers with limited bioinformatics expertise. Therefore, Galaxy has been chosen for NGS data analysis, particularly for users without extensive command-line expertise. Different Galaxy workflows have been built to analyze RNA-Seq Data [21], *de novo* bacterial genome assembly and annotation [22], 16s meta-genomic analysis [23]. The popularity of Galaxy platform in analyzing NGS sequencing data has surged due to its user-friendly interface, workflow reproducibility, integrated tools, community support, and adaptability to cloud computing, making it accessible and effective for researchers across diverse backgrounds, such as physicians or medical doctors in a clinical genetics laboratory setting [24].

4. CONCLUSIONS

In this study, we conducted a comparative analysis of exome sequencing data from syndactyly abnormalities using the two platforms: GALAXY and UNIX. Our findings revealed notable differences in variant calling and shared variants between the two platforms. The UNIX platform identified a total of 275,572 variants, comprising 248,392 SNPs and 27,180 indels, while the GALAXY platform identified 140,291 variants, including 129,792 SNPs and 10,499 indels, from the exome data. A common set of 126,848 variants, with 117,339 SNPs and 9,509 indels, was identified between the two pipelines. Subsequent filtration based on a list of 200 genes with the highest syndactyly-related scores resulted in 1,345 retained variants, encompassing 1,257 SNPs and 88 indels. These variants are distributed throughout the patient's genome, concentrating on certain chromosomes such as chromosomes 2, 4, and 11. Meanwhile, the genes containing the highest number of variants within the top 200 genes associated with syndactyly were *FRAS1*, *CACNA1C*, *GLI2*, and *NOTCH1*. This study emphasized the impact of bioinformatic tools on the identification of genetic variations associated with syndactyly. For users seeking a more accessible and reproducible platform with a user-friendly interface, Galaxy is recommended, particularly for smaller datasets or routine analyses. On the other hand, Unix is preferred for users requiring higher performance, flexibility, and the ability to handle large-scale data and complex custom analyses. . Our results provide valuable insights into the selection of analytical platforms for exome sequencing data in the context of congenital limb abnormalities, contributing to the optimization of variant detection workflows.

Acknowledgements. This study received financial support from the project with code VAST01.04/23-24, sponsored by the Vietnam Academy of Science and Technology.

Authorship contribution statement. Nguyen Thy Ngoc: Methodology, Investigation, Formal analysis, Manuscript writing, Funding acquisition. Huynh Minh Huong: Formal analysis.

Declaration of competing interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Ahmed H., Akbari H., Emami A., Akbari M. R. - Genetic Overview of Syndactyly and Polydactyly. Plastic and reconstructive surgery, Global open **5** (2017) e1549.
2. Mandal K., Phadke S. R., Kalita J. - Congenital swan neck deformity of fingers with syndactyly, Clinical dysmorphology **17** (2008) 109-111.
3. Malik S., Afzal M., Gul S., Wahab A., Ahmad M. - Autosomal dominant syndrome of camptodactyly, clinodactyly, syndactyly, and bifid toes. American journal of medical genetics, Part A **152A** (2010) 2313-2317.

4. Vieira C., Teixeira N., Cadilhe A., Reis I. - Apert syndrome: prenatal diagnosis challenge, BMJ case reports **12** (2019)
5. Turnpenny P. D., Dean J. C., Duffty P., Reid J. A., Carter P. - Weyers' ulnar ray/oligodactyly syndrome and the association of midline malformations with ulnar ray defects, Journal of medical genetics **29** (1992) 659-662.
6. Patel R., Singh S. K., Bhattacharya V., Ali A. - Novel HOXD13 variants in syndactyly type 1b and type 1c, and a new spectrum of TP63-related disorders, Journal of human genetics **67** (2022) 43-49.
7. Ngoc N. T., Duong N. T., Quynh D. H., Ton N. D., Duc H. H., Huong L. T. M., Anh L. T. L., Hai N. V. - Identification of novel missense mutations associated with non-syndromic syndactyly in two vietnamese trios by whole exome sequencing, Clinica chimica acta, International Journal of Clinical Chemistry **506** (2020) 16-21.
8. Deng H., Tan T. - Advances in the Molecular Genetics of Non-syndromic Syndactyly, Current genomics **16** (2015) 183-193.
9. Jelin A. C., Vora N. - Whole Exome Sequencing: Applications in Prenatal Genetics, Obstetrics and gynecology clinics of North America **45** (2018) 69-81.
10. Blankenberg D., Gordon A., Von Kuster G., Coraor N., Taylor J., Nekrutenko A., Galaxy T. - Manipulation of FASTQ data with Galaxy, Bioinformatics **26** (2010) 1783-1785.
11. Van der Auwera G. A., Carneiro M. O., Hartl C., Poplin R., Del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K. V., Altshuler D., Gabriel S., DePristo M. A. - From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, Current protocols in bioinformatics **43** (2013) 11 10 11-11 10 33.
12. Al-Qattan M. M. - A Review of the Genetics and Pathogenesis of Syndactyly in Humans and Experimental Animals: A 3-Step Pathway of Pathogenesis, BioMed research international **2019** (2019) 9652649.
13. Cassim A., Hettiarachchi D., Dissanayake V. H. W. - Genetic determinants of syndactyly: perspectives on pathogenesis and diagnosis, Orphanet journal of rare diseases **17** (2022) 198.
14. Hines E. A., Verheyden J. M., Lashua A. J., Larson S. C., Branchfield K., Domyan E. T., Gao J., Harvey J. F., Herriges J. C., Hu L., McCulley D. J., Throckmorton K., Yokoyama S., Ikeda A., Xu G., Sun X. - Syndactyly in a novel Fras1(rdf) mutant results from interruption of signals for interdigital apoptosis, Developmental dynamics: an official publication of the American Association of Anatomists **245** (2016) 497-507.
15. Chen X., Birey F., Li M. Y., Revah O., Levy R., Thete M. V., Reis N., Kaganovsky K., Onesto M., Sakai N., Hudacova Z., Hao J., Meng X., Nishino S., Huguenard J., Pasca S. P. - Antisense oligonucleotide therapeutic approach for Timothy syndrome, Nature **628** (2024) 818-825.
16. Pan Y., Liu Z., Shen J., Kopan R. - Notch1 and 2 cooperate in limb ectoderm to receive an early Jagged2 signal regulating interdigital apoptosis, Developmental biology **286** (2005) 472-482.
17. Minhas R., Pauls S., Ali S., Doglio L., Khan M. R., Elgar G., Abbasi A. A. - Cis-regulatory control of human GLI2 expression in the developing neural tube and limb bud. Developmental dynamics: an official publication of the American Association of Anatomists **244** (2015) 681-692.

18. Koboldt D. C. - Best practices for variant calling in clinical sequencing, Genome medicine **12** (2020) 91.
19. Afgan E., Baker D., Batut B., van den Beek M., Bouvier D., Cech M., Chilton J., Clements D., Coraor N., Gruning B. A., Guerler A., Hillman-Jackson J., Hiltemann S., Jalili V., Rasche H., Soranzo N., Goecks J., Taylor J., Nekrutenko A., Blankenberg D. - The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic acids research **46** (2018) W537-W544.
20. Sims D., Ilott N. E., Sansom S. N., Sudbery I. M., Johnson J. S., Fawcett K. A., Berlanga-Taylor A. J., Luna-Valero S., Ponting C. P., Heger A. - CGAT: computational genomics analysis toolkit, Bioinformatics **30** (2014) 1290-1291.
21. Batut B., van den Beek M., Doyle M. A., Soranzo N. - RNA-Seq Data Analysis in Galaxy, Methods in molecular biology **2284** (2021) 367-392.
22. Wee S. K., Yap E. P. H. - GALAXY Workflow for Bacterial Next-Generation Sequencing De Novo Assembly and Annotation, Current protocols **1** (2021) e242.
23. Thang M. W. C., Chua X. Y., Price G., Gorse D., Field M. A. - MetaDEGalaxy: Galaxy workflow for differential abundance analysis of 16s metagenomic data, F1000Research **8** (2019) 726.
24. Chappell K., Francou B., Habib C., Huby T., Leoni M., Cottin A., Nadal F., Adnet E., Paoli E., Oliveira C., Verstuyft C., Davit-Spraul A., Gaignard P., Lebigot E., Duclos-Vallee J. C., Young J., Kamenicky P., Adams D., Echaniz-Laguna A., Gonzales E., Bouvattier C., Linglart A., Picard V., Bergoin E., Jacquemin E., Guiochon-Mantel A., Proust A., Bouligand J. - Galaxy Is a Suitable Bioinformatics Platform for the Molecular Diagnosis of Human Genetic Disorders Using High-Throughput Sequencing Data Analysis: Five Years of Experience in a Clinical Laboratory, Clinical chemistry **68** (2022) 313-321.