Genetic Characterization of Thai Indigo (*Indigofera tinctoria*) by Data Single-Nucleotide Variants (SNVs) analysis

Prarak Amornsak[1], Nawatthakorn Umasin[2],
Wasan Palasai [3] and Sulaiman Cheabu[4*]

Received: August 5, 2025; Revised: August 26, 2025;
Accepted: November 3, 2025; Published: November 20, 2025;

Abstract

In the past, southern indigo was a highly sought-after local product for Portuguese traders, but the traditional knowledge and practices surrounding it have gradually faded over time. Meanwhile, in Northeastern Thailand, various types of indigo are widely cultivated, with *Indigofera tinctoria L.* (dye indigo) and *Indigofera suffruticosa* (wild indigo) being popular for producing a deep blue dye. Beyond its use as a dye, the indigo plant is also a source of bioactive compounds. This study aimed to investigate the genetic diversity of three indigo varieties: "Kram Thale" (sea indigo), "Kram Fak Khong" (curved-pod indigo), and "Kram Fak Trong" (straight-pod indigo). The research utilized single-nucleotide variants (SNVs) as molecular markers and analyzed the chemical composition of key compounds. The results showed that the nucleotide sequence data were of high quality, with an average quality score of 37–38. A total of 2,637,721 SNVs were initially detected, which was reduced to 1,801,689 SNVs after filtering. Genetic relationship analysis using a phylogenetic tree and Principal Component Analysis (PCA) revealed two main clusters. "Kram Thale" and "Kram Fak Khong" were found to be genetically similar, while "Kram Fak Trong" was distinctly separate. Population structure analysis also clearly divided the samples into three groups based on their variety. Furthermore, chemical analysis of the "Kram Thale" leaf extract using GC-MS identified several important compounds, including Phytol (6.73%), Hexadecanoic acid, 2-hydroxy-1-(hydroxymethyl) ethyl- ester (3.71%), 7-9-di+tert butyl-1-oxaspiro (4,5) deca-6,9-dinen-2,8-dione (3.62%), and 4-Acetyl-1,2,3,4-tetrahydro-2-oxoquinoline (3.41%). This study highlights the genetic differences between indigo varieties and the significant chemical composition of "Kram Thale," providing a foundation for future breeding programs and product development.

Keywords: Indigo, *Indigofera tinctoria*, single-nucleotide variants (SNVs), genetic diversity, bioactive compounds

[1].Faculty of management science, Princess of Naradhiwas University, Naradhiwas 96000, Thailand

[2] Faculty of architecture, Rajamangala University of Technology Srivijaya, Songkhla 90000, Thailand

[3] Department of Mechanical Engineering, Faculty of Engineering, Princess of Naradhiwas University, Narathiwat 96000, Thailand

[4] Faculty of Agriculture, Princess of Naradhiwas University, Naradhiwas 96000, Thailand

* Corresponding Author. E-mail: sulaiman.cheabu@gmail.com

Introduction

Indigofera tinctoria L., a plant belonging to the family Fabaceae, is native to Southeast Asia and is widely cultivated in Asia, Africa, and Central America. It is a plant of significant economic and cultural importance, particularly for producing natural indigo dye, which has been used for centuries to color cotton and other natural fibers. In addition to its use as a dye, indigo is also used in traditional medicine to treat various ailments, such as skin, liver, and respiratory diseases. In Thailand, indigo has been cultivated for a long time, especially in the northern and northeastern regions, where it is used to dye local textiles like cotton, silk, and "pha sin". The demand for naturally dyed products and local wisdom has led to a growing popularity of indigo-dyed fabrics and the promotion of indigo cultivation in many parts of Thailand. The classification of indigo species in Thailand is traditionally based on morphological characteristics like leaves, flowers, and pods, which has led to the identification of several types such as "Kram Yai" (large indigo), "Kram Lek" (small indigo), "Kram Fak Trong" (straight-pod indigo), "Kram Fak Ngor" (curved-pod indigo), and "Kram Thale" (sea indigo). However, relying solely on morphological traits can be limiting, as these characteristics can be influenced by environmental factors, leading to inaccurate classification.

Whole-genome sequencing (WGS) is a powerful tool for studying the genetic diversity of plants. The use of single-nucleotide variants (SNVs), which are changes at a single nucleotide position, is particularly effective because they are found throughout the genome, allowing for accurate genetic relationship studies and variety classification (Chedao, N.,2024). Furthermore, studying the chemical composition of plants is crucial for identifying bioactive compounds with potential medical and pharmaceutical applications. While previous research has examined the genetic characteristics of indigo in other countries, such as India, China, and Indonesia, the genetic makeup of Thai indigo has been studied to a limited extent, especially using WGS combined with SNV analysis. Therefore, this research aims to investigate the genetic diversity of three Thai indigo varieties—"Kram Thale," "Kram Fak Khong," and "Kram Fak Trong"—by using SNV molecular markers and analyzing the chemical composition of key compounds. The findings will provide a fundamental basis for the future conservation, breeding, and product development of indigo.

Methodology

1. Sample Preparation

" (B1, B2), and "Kram Fak Trong" (S1, S2), for a total of six samples. Healthy leaves free from signs of disease or pest infestation were selected. The samples were then stored at 4°C to maintain their condition. DNA was extracted from all six leaf samples using a DNeasy Plant Mini Kit (Qiagen, Germany). The fresh leaves were first ground to a fine powder in liquid nitrogen, and the powder was placed in a plant cell lysis buffer with RNase A enzyme. The mixture was incubated at 65°C for 10 minutes, and a protein precipitation buffer was added before the protein pellet was removed by centrifugation. The supernatant was then transferred to a column and centrifuged to precipitate the DNA. The quality and concentration of the extracted DNA were measured using a Nanodrop Spectrophotometer and a Qubit Fluorometer (Thermo Fisher Scientific, USA). The integrity of the DNA was also checked by gel electrophoresis on a 1% agarose gel. DNA samples with an A260/A280 ratio between 1.8 and 2.0 were selected for subsequent analysis

2. Whole-Genome Sequencing

Paired-end DNA libraries were constructed using DNA nanoball sequencing. The target DNA was

fragmented into pieces approximately 350 bp in size, and adapters were ligated to the fragments. The DNA was then amplified using PCR, and whole-genome sequencing was performed on a DNBSEQ-G400 instrument (BGI, China) with a read length of 150 bp. The target average coverage was set to 10x per sample. Data quality was checked using the FastQC v0.11.9 program (Andrews, 2010) by evaluating the mean quality scores across the read length (0–150 bp). Reads were then filtered using Trimmomatic v0.39 (Bolger et al., 2014). SNVs were identified by aligning the filtered reads to a reference genome using the Bowtie 2 v2.4.2 program (Langmead & Salzberg, 2012) with default parameters. The alignment files were refined using Picard Tools v2.25.0. High-quality SNVs for genetic diversity analysis were selected using a minor allele frequency of 2% with the VCFtools v0.1.16 program (Danecek et al., 2011). Finally, a Phylogenetic Tree, Principal Component Analysis (PCA), and Population Structure analysis were performed to study the genetic diversity of "Kram Thale" and "Kram Fak Ngor" using "Kram Fak Trong" as a control.

3. Analysis of Bioactive Compounds

Fresh "Kram Thale" leaves were collected and washed with distilled water. The leaves were then dried in a hot air oven at 45°C for 48 hours and ground into a fine powder with an electric grinder. A 50-gram sample of the powdered leaves was soaked in 250 ml of 95% ethanol for 72 hours at room temperature, with occasional shaking. The mixture was filtered through Whatman No. 1 filter paper, and the solvent was evaporated from the filtrate using a vacuum evaporator at 45°C to obtain a crude extract. The crude extract was stored in a freezer at -20°C. The chemical composition of the crude extract from the "Kram Thale" leaves was analyzed using Gas Chromatography-Mass Spectrometry (GC-MS) at the Scientific Instrument Center, Prince of Songkhla University, Hat Yai Campus. An Agilent 7890B GC coupled with an Agilent 5977A MS detector and an HP-5MS column (30 m × 0.25 mm, film thickness 0.25 μm) was used. Compounds were identified by comparing their mass spectra with the NIST (National Institute of Standards and Technology) Mass Spectral Library and the Wiley Registry of Mass Spectral Data. Compounds with a similarity value greater than 80% were considered, and the percentage of each compound was calculated by peak area normalization.

Results and discussions

1.Quality Control of Whole-Genome Sequencing (WGS) Data

DNA was extracted from three indigo samples using a DNeasy Plant Mini Kit to obtain high-quality DNA for WGS. Quality analysis of the paired-end DNA nanoball sequencing data using FastQC (Andrews, 2010) showed that the mean quality scores ranged from 37 to 38 across the entire read length (0–150 bp). These scores indicate very high and reliable read quality, with consistent scores suggesting a stable sequencing process. The data is \therefore of high quality and suitable for further analysis (Figure 1)

.

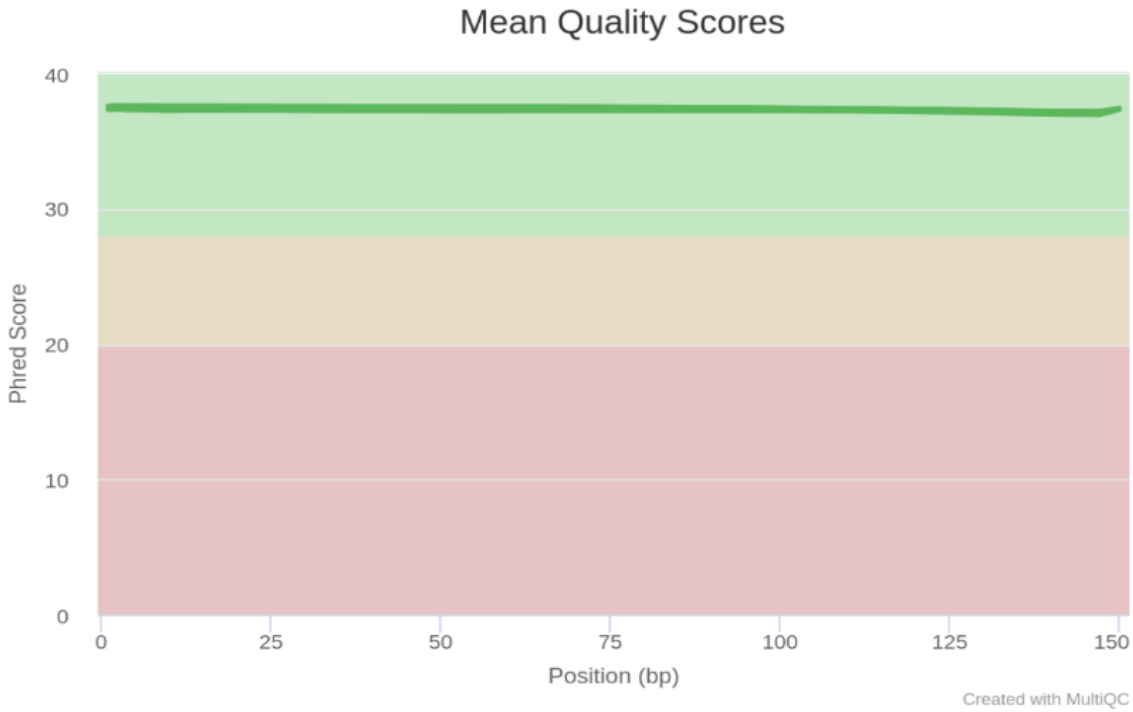


Figure 1. FastQC mean quality scores. FastQC quality scores for all six samples were obtained. The higher the phred score, the better the base call. For all six samples, bases for all samples were considered high quality (green). In addition, the quality scores remained consistent across the entire read lengt

2. Alignment with Genome Assembly and SNV Discovery

Genome assembly data with contig lengths of ≥ 5000 bp were used as a reference genome for SNV discovery. Reads were aligned to the reference genome using Bowtie 2 (Langmead & Salzberg, 2012). The alignment results showed an average mapped read percentage of 40.24%, an average read depth of 11.01, and an average genome coverage of 92.45%. SNVs and Indels were discovered using the GATK (Genome Analysis Toolkit) suite (Van der Auwera et al., 2013), with a total of 3,073,703 variants found. This included 2,637,721 SNVs and 435,982 small-Indels. These SNVs were then filtered to select high-quality variants for genetic diversity analysis, using a minor allele frequency of 2% and excluding positions with N bases or missing data. This process yielded 1,801,689 high-quality SNVs (Table 1)

Table 1. Alignment results of samples with the reference genome

Sample ID	Sequencing ID	Mapped read (M)	Percent of mapped read (%)	Read depth average	Genome coverage (%)
InT1	1	19.99	43.30%	17.6564	96.26%
InT2	2	11.27	43.39%	10.2454	92.82%
InB1	3	10.15	42.28%	9.3869	91.69%
InB2	4	10.30	39.81%	9.18672	92.01%
InS1	5	11.71	36.53%	11.193	91.77%
InS2	6	8.83	36.15%	8.40423	90.13%
	Average	12.04	40.24%	11.01	92.45%
	Min	8.83	36.15%	8.40	90.13%
	Max	19.99	43.39%	17.66	96.26%

3. Genetic Diversity Analysis Using Phylogenetic Tree, Principal Component Analysis (PCA), and Population Structure

The 1,801,689 high-quality SNVs were used for phylogenetic tree analysis using the Neighbor-Joining method in MEGA X (Kumar et al., 2018). The PCA analysis was performed in R using the SNPRelate package for dissimilarity calculations and the ggplot2 package for plotting. Population structure analysis from the SNVs was performed using the ADMIXTURE program (Alexander et al., 2009). The analysis of the relationship among the six samples from three varieties "Kram Thale" (T1, T2), "Kram Fak Ngor" (B1, B2), and "Kram Fak Trong" (S1, S2) showed that the phylogenetic tree analysis, with a standard distance of 0.1, clearly divided the samples into two main groups. The first group consisted of "Kram Thale" (T1, T2) and "Kram Fak Ngor" (B1, B2), which were closely related, while "Kram Fak Trong" (S1, S2) formed a separate, significantly different group (Figure 2a).

Principal Component Analysis (PCA) showed the distribution of the samples in a two-dimensional space, with PC1 explaining 96.85% of the variance and PC2 explaining 1.84%. The PCA results confirmed the grouping pattern observed in the phylogenetic tree. "Kram Fak Trong" (S1, S2) was in the upper right quadrant of the plot, while "Kram Fak Ngor" (B1, B2) was in the upper left, and "Kram Thale" (T1, T2) was in the lower left, indicating significant differences between the groups. Furthermore, the clustering of the three indigo varieties based on relative abundance clearly divided the samples into three distinct groups. Group 1 (red) consisted of "Kram Thale" (T1, T2), Group 2 (green) of "Kram Fak Trong" (S1, S2), and Group 3 (blue) of "Kram Fak Ngor" (B1, B2). This clustering pattern is consistent with the results from the phylogenetic tree and PCA analyses.

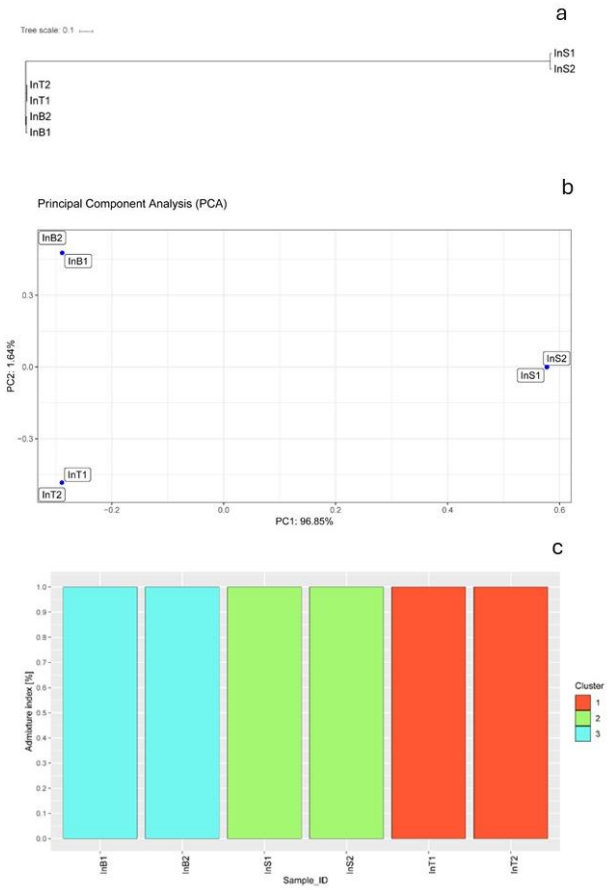


Figure 2: a) Phylogenetic tree using the Neighbor-Joining method, b) Grouping results by PCA, and c) Population structure analysis dividing samples into three groups for six samples from three indigo varieties: Kram Thale (T1, T2), Kram Fak Ngor (B1, B2), and Kram Fak Trong (S1, S2)

4. Identification of Bioactive Compounds By GC-MS

The crude extract from "Kram Thale" leaves was analyzed using GC-MS at the Scientific Instrument Center, Prince of Songkhla University, Hat Yai Campus. The analysis of the main chemical components within the crude extract from the leaves identified several key compounds. These included Phytol, Hexadecanoic acid, 2-hydroxy-1-(hydroxy-methyl) ethyl ester, 7-9-di+tert butyl-1-oxaspiro (4,5) deca-6,9-dinen-2,8-dione, and 4-Acetyl-1,2,3,4-tetrahydro-2-oxoquinoline, with percentages of 6.73%, 3.71%, 3.62%, and 3.41%, respectively (Table 2 and Figure3)

Table 2 Chemical analysis results of the main components within the crude indigo leaf extract

Compound Name	% of total
Phytol	6.73%
Hexadecanoic acid, 2-hydroxy-1-(hydroxymethyl)ethyl ester	3.71%
7-9-di+tert butyl-1-oxaspiro (4,5) deca-6,9-dinen-2,8-dione	3.62%
4 Acetyl-1,2,3,4-tetrahydro-2-oxoquinoline	3.41 %

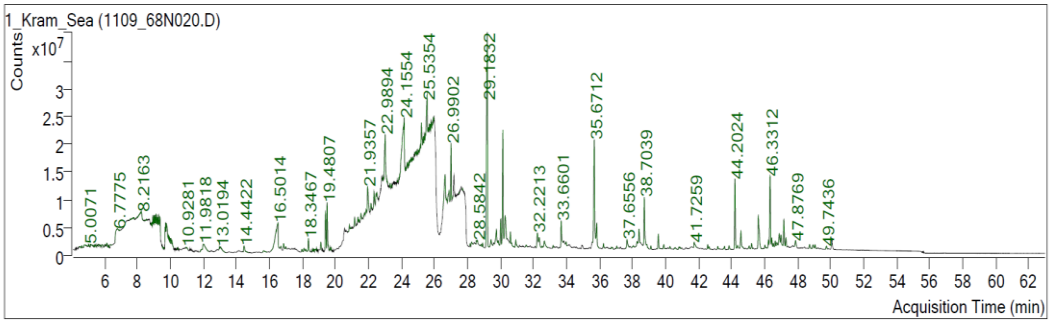


Figure 3 Chromatogram of crude extract from "Kram Thale" leaves analyzed by GC-MS

Conclusion

This study on the genetic characterization of *Indigofera tinctoria* using SNV molecular markers and chemical composition analysis yielded several key findings. The whole-genome sequencing using DNA nanoball sequencing provided high-quality data, with mean quality scores ranging from 37 to 38, which remained consistent throughout the reads. Alignment of the reads to a reference genome showed an average of 40.24% alignment with a mean depth of 11.01, covering 92.45% of the reference genome. A total of 2,637,721 SNVs were detected, and 1,801,689 high-quality SNVs were retained after filtering.

Genetic relationship analysis using a phylogenetic tree showed that the samples were divided into two main groups, with "Kram Thale" and "Kram Fak Ngor" being closely related, while "Kram Fak Trong" was clearly separate. The PCA analysis confirmed this, with PC1 explaining 96.85% of the variance, indicating clear genetic differences between the sample groups. Population structure analysis also showed a distinct separation into three groups corresponding to the three varieties: "Kram Thale," "Kram Fak Ngor," and "Kram Fak Trong". Chemical analysis of the "Kram Thale" leaf extract by GC-MS identified several important compounds, notably Phytol (6.73%), which is known for its antioxidant and anti-inflammatory properties. This study demonstrates significant genetic differences among the three indigo varieties and the presence of important bioactive compounds in "Kram Thale". These findings have the potential to be further developed into medical or pharmaceutical products in the future.

References

Alexander, D. H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655-1664.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17(2), 81-92.

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

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

Chanayath, N., Lhieochaiphant, S., Phutrakul, S. (2002). Pigment extraction techniques from the leaves of Indigofera tinctoria Linn. and Baphicacanthus cusia Brem. and chemical structure analysis of their major components. Chiang Mai University Journal of Natural Sciences, 1(2), 149-160.

Chedao, N. (2024). Evaluation of Genetic Diversity of Tungat Ali (Eurycoma sp.) Using Simple Sequence Repeat (SSR) Marker. *International Journal of Science and Innovative Technology*, *7*(2), 44-52.

Chutichudet, P., Chutichudet, B., Boontiang, K. (2015). Diversity of Indigo plants (*Indigofera tinctoria* L.) in Thailand. International Journal of Agricultural Technology, 11(8), 1691-1699.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution, 29(1), 51-63.

Gilbert, K. G., Maule, H. G., Rudolph, B., Lewis, M., Vandenburg, H., Sales, E., Tozzi, S., Cooke, D. T. (2002). Quantitative analysis of indigo and indigo precursors in leaves of Isatis spp. and Polygonum tinctorium. Biotechnology Progress, 18(6), 1214-1218.

Gulsen, O., Uzun, A., Canan, I., Seday, U., Canihos, E. (2010). A new citrus linkage map based on SRAP, SSR, ISSR, POGP, RGA and RAPD markers. Euphytica, 173(2), 265-277.

Han, R., Takahashi, H., Nakamura, M., Bunsupa, S., Yoshimoto, N., Yamamoto, H., Suzuki, H., Shibata, D., Yamazaki, M., Saito, K. (2019). Transcriptome analysis of nine tissues to discover genes involved in the biosynthesis of active ingredients in Sophora flavescens. Biological and Pharmaceutical Bulletin, 42(7), 1185-1194.

Kumar, S., Bhardwaj, T. R., Prasad, D. N., Singh, R. K. (2020). Natural products: A treasure for therapeutic candidates. World Journal of Pharmacy and Pharmaceutical Sciences, 9(3), 1405-1414.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution, 35(6), 1547-1549.

Laitonjam, W. S., Wangkheirakpam, S. D. (2011). Comparative study of the major components of the indigo dye obtained from Strobilanthes flaccidifolius Nees. and Indigofera tinctoria Linn. International Journal of Plant Physiology and Biochemistry, 3(7), 108-116.

Langmead, B., Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357-359.

Muzayyinah, M. (2020). Genetic diversity of Indonesian natural indigo-producing plant revealed by RAPD markers. Biodiversitas Journal of Biological Diversity, 21(4), 1578-1583.

Prathomrach, W., Padungkul, P., Ratanatriwong, P. (2018). Natural indigo dyeing process on cotton fabric by using fermentation techniques from Thailand. Journal of Applied Arts, 11(1), 111-121.

Singh, R., Jain, A., Panwar, S., Gupta, D., Khare, S. K. (2011). Antimicrobial activity of some natural dyes. Dyes and Pigments, 66(2), 99-102.

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. (2013). From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics, 43(1), 11.10.1-11.10.33.

Verma, S. M., Singh, L. (2014). Therapeutic dimensions of Indigofera tinctoria Linn: An overview. International Journal of Pharmaceutical Sciences Review and Research, 28(1), 169-173.