BMC Research Notes

DATA NOTE

Open Access

De novo transcriptome assembly of *Clitoria ternatea* flowers

Nallammai Singaram[1,4], Muhamad Afiq Bin Mohamed Akbar[2], Hamidun Bunawan[3] and Wei Hsum Yap[1,4*]

Abstract

Objective The *Clitoria ternate* flowers are widely used for its anthocyanin pigmentation. However, the flowers' colour ranges from white to light blue to deep blue, purple or even pink. The butterfly pea flower colours are known to be influenced by variety and not environmental factors. Despite its interest there is a lack of understanding of the level of gene expression between varieties, in producing blue and white butterfly pea flowers. Hence, transcriptome assembly of genes expressed in different butterfly pea flowers and subsequent understanding of the key genes will help in increasing the pigments accordingly.

Data description An RNA library was generated from RNA extracted from three different sources of flower petals as in white, light blue and dark blue butterfly pea flowers and sequenced using the Illumina HiSeq Platform. These data sets were then used to produce and assemble the first transcriptome assembly based on *Clitoria ternate* or commonly known as butterfly pea, paving way for anthocyanin related gene identification.

Keywords RNA-seq, Transcriptome, De novo assembly, *Clitoria ternate*, Butterfly pea

Objective

Clitoria ternatea L. commonly known as the butterfly pea, is widely cultivated in Southeast Asia as both an ornamental and a medicinal plant. Beyond its traditional use in food and beverages, the species is valued for its rich phytochemical composition, particularly anthocyanins, which function as natural colorants and antioxidants [1, 2]. Anthocyanins are water-soluble flavonoid pigments responsible for a diverse range of colors including orange, red, purple, violet, and blue in plant tissues. Their role besides as pollinator attraction is to provide UV protection in plants, and promote health in humans, such as antibacterial, hypoglycemic, and antiproliferative effects [3–7].

In *C. ternatea*, anthocyanins known as ternatins impart the vivid blue coloration of the flowers [8, 9]. However, the regulatory elements and transcriptomic framework underlying ternatin biosynthesis remain poorly characterized. The absence of comprehensive genomic information has limited deeper exploration of the anthocyanin biosynthetic pathway in this species, despite its emerging significance in agriculture, food technology, and pharmacology.

Clitoria ternatea flower is one anthocyanin source containing stable blue colour polyacylated anthocyanins [10, 11]. The presence of polyacylated anthocyanins, metal ions, other phenolic compounds and the resulting co-pigmentation effect may assist to form more stable and

*Correspondence:
Wei Hsum Yap
weihsum.yap@taylors.edu.my
[1]School of Biosciences, Faculty of Health and Medical Sciences, Taylor's University, Jalan Taylor's, Subang Jaya, Selangor, Malaysia
[2]Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43000 Seri Kembangan, Selangor, Malaysia
[3]Institute of Systems Biology (INBIOSIS), Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia
[4]Food Security and Nutrition Impact Lab, Taylor's University, No. 1, Jalan Taylor's, 47500 Subang Jaya, Selangor, Malaysia

intense blue colours [12, 13]. Anthocyanin biosynthesis involves a well-established pathway of structural and regulatory genes whose expression is strongly influenced by environmental conditions such as light, temperature, and nutrient availability [14]. These regulatory dynamics likely contribute to the variation in flower pigmentation observed among different *C. ternatea* varieties, particularly white, pale blue, and dark blue flowers. Yet, systematic investigations comparing gene expression, metabolite accumulation, and tissue localization across these color variants are lacking.

As a first step towards the goal of spatial validation of pigment deposition that is tied to morphology and expression would be transcriptome analysis based on expression in blue, white and light-blue or sky-blue butterfly-peas. In this paper, a brief analysis of the transcriptome data of three variants of butterfly-peas based on colour is being presented. Here we report the assembled and annotated transcriptome of three different variants of butterfly pea based on flower petal RNA extracts. The analysis will serve as valuable genomic information for subsequent metabolic association and gene identification for future applications in sustainable agriculture and plant biotechnology.

Flower samples and RNA quality

Commercial *Clitoria ternatea* L. seeds (SHS Kebun, Kulim, Malaysia) of dark blue, light/sky blue, and white flower variants were cultivated in plastic pots (30 cm in diameter) containing commercial potting soil mix at Taylor's University, Selangor, Malaysia under natural photoperiod (about 12 h) and ambient temperature conditions (28 ± 2 ℃). A total of 18 plants were grown from seeds to flowering stage, two pots per variant, and 3 plants per pot. The plants were adequately provided with water, NPK (10:10:10) fertilizers, and regular pruning. The second round of blooms after ensuring flower colour was used to avoid metabolic fluctuations. Flower samples used are fully opened fresh new flowers harvested between 11 a.m. to 12 noon to minimize circadian variation in gene expression throughout the research and were immediately transported to the laboratory on ice. The samples were prepared, and RNA was extracted following Fareed and Singaram [15] method of RNA extraction using NucleoSpin® RNA Plant kit (Macherey-Nagel, Germany). The purified RNA was assessed and subsequently sent to Azenta SZ NGS Laboratory (Bio3 Sdn. Bhd.) for sequencing.

High-quality total RNA was successfully extracted from all butterfly pea flower samples (Table 1, Data File 1 [16]). RNA concentrations ranged from 80.3 to 390.9 ng/µl, yielding total RNA amounts between 3.29 µg and 13.9 µg per sample, which reflects inherent differences in tissue type, developmental stage, or physiological state at the

Table 1 Overview of data files and data sets

Label	Name of data file/ data set	File type	Data repository	Data repository and identifier; References
Data file 1	Quality and Quantity of extracted total RNA and the RIN value	Image (.jpg) FASTQ (fastq)	Figshare	https://doi.org/10.6084/m9.figshare.30487016 [19]
Data file 2	RNA-seq dataset (SRA) obtained in this study		NCBI Sequence Read Archive (SRA)	https://identifiers.org/ncbi/insdc.sra:SRP627045 [20]
Data file 3	Methodology	MS Word (.docx)	Figshare	https://doi.org/10.6084/m9.figshare.30486929 [21]
Data file 4	MultiQC read quality results	MS Word (.docx)	Figshare	https://doi.org/10.6084/m9.figshare.30472619 [22]
Data file 5	RNA-seq de novo transcriptome assembly	FASTA (.fasta)	Figshare	https://doi.org/10.6084/m9.figshare.30472562 [23]
Data file 6	*Clitoria ternatea* flower transcriptome shotgun assembly	FASTA (.fasta)	GenBank TSA	https://identifiers.org/ncbi/insdc:GLLN00000000.1 [24]
Data file 7	Protein sequence prediction	FASTA (.fasta)	Figshare	https://doi.org/10.6084/m9.figshare.30472592 [25]
Data file 8	Gene Ontology annotation	Image (.jpg)	Figshare	https://doi.org/10.6084/m9.figshare.30465206 [26]
Data file 9	Functional annotation	MS Excel (.xlsx)	Figshare	https://doi.org/10.6084/m9.figshare.30486899.v2 [27]

time of sampling [17], even though strict measures were taken to include coloured portions of the petals, first day full-bloom harvest, and consistent harvesting time of the day. Purity ratios of A260/280 ranged 2.15–2.23, and A260/230 was between 1.65 and 2.18, this confirmed minimal protein and solvent contamination, indicating that the extracted RNA was of sufficient purity for downstream sequencing applications. Furthermore, the RNA Integrity Number (RIN) values were consistently high, ranging from 7.1 to 10.0, with majority of the samples scoring above 9.0. Since RIN values ≥ 7 is considered suitable for RNA-Seq library preparation, all samples met the required quality threshold [15–18]. The samples PBS3, SBS1, SBS2, SBS3, and PWS3 achieved the maximum RIN score of 10.0, reflecting excellent RNA integrity, while PBR2 (RIN 7.1) and PWR4 (RIN 8.1) still fell within acceptable ranges. Overall, these results confirmed that the extracted RNA was of high integrity, purity, and yield, validating its suitability for Illumina RNA-Seq library construction and comparative transcriptomic analyses.

Data description

To generate a whole-plant library for each flower colour as in dark blue (PB), white (PW), and sky-blue (SB), three replicates from independent plants per colour variant were used. A standard amount of intact, high-quality total RNA (RIN value > 7.0) for each biological replicate of each colour variety was used for cDNA library construction and sequenced on the Illumina NovaSeq X Plus platform (GENEWIZ Biotechnology Co. Ltd., Beijing, China). The Raw Illumina RNA-seq data were deposited in the NCBI Sequence Read Archive (SRA) database SRP627045 (Table 1, Data File 2 [19]). The Transcriptome de novo assembly and annotation methodology are reported in Data File 3 (Table 1) [20].

Raw RNA-Seq reads were subjected to quality control and trimming using Trimmomatic preprocessing tools, and the resulting clean reads were assembled de novo with Trinity, applying a minimum k-mer coverage of 2 to minimize mis-assemblies. Nine RNA-Seq libraries were generated from three variants of butterfly pea (*Clitoria ternatea*). Read quality was evaluated using FASTQC (v0.12.1) and summarized with MultiQC (v1.14). Each RNA-Seq library produced an average of 66.4 to 82.5 million paired-end reads with average read lengths of about 146 bp. Quality analysis showed that ≥ 97% and ≥ 93% of bases had Phred quality scores above Q20 and Q30, respectively, with a mean per-base error rate of < 0.1%. The GC content ranged from 42.8% to 43.3%, and the frequency of ambiguous bases (N) remained < 3.7 ppm across all libraries, suggesting minimal sequencing artifacts. Thus, no significant adapter contamination or GC bias was observed, confirming high sequencing accuracy and suitability for downstream analyses [21], Data File 4 (Table 1) [22] illustrates the distribution of the mean read counts per quality score for both first-in-pair and second-in-pair reads. Across all sequencing cycles, the first-in-pair reads maintained consistently high base-calling accuracy, with average Phred quality scores exceeding 33, indicating an error probability of less than 0.0005 per base. The second-in-pair reads also displayed similarly high-quality profiles, with only a slight decline in the terminal bases, a trend commonly observed in Illumina sequencing runs. Overall, these results demonstrate the robustness of library preparation and the suitability of the dataset for reliable downstream transcriptomic analyses.

The longest isoform from each gene cluster was selected to represent unigenes for downstream analyses. The filtered reads were assembled into 106,344 unigenes with an average length of 889.56 bp and N50 of 2022 bp using the de novo assembler Trinity v2.15.1. The RNA-seq De novo transcriptome assembly dataset (Table 1, Data file 5 [23]), has been deposited in the Transcriptome Shotgun Assembly database of NCBI under accession number GLLN00000000 (Table 1, Data file 6 [24]). A total of 455,232 open reading frames (ORFs) were predicted (Table 1, Data file 7 [25]).

The Functional annotation of unigenes through sequence similarity searches against the NCBI non-redundant protein (Nr) (49098 hits), InterPro (27713 hits), KEGG (13688 hits), eggnog (28472 hits), and Gene Ontology (GO) databases (35210 hits) was performed. GO terms were further classified and visualized using WEGO 2.0 to illustrate gene distribution across biological processes, molecular function, and cellular component categories.

Gene Ontology (GO) annotation of the assembled unigenes was performed using Blast2GO, categorizing them into three major GO domains: Cellular Component, Molecular Function, and Biological Process (Table 1, Data File 8 [26]). A total of 106,344 genes were successfully annotated, representing diverse cellular roles and molecular activities. Within the Cellular Component category, majority of genes were associated with the cell, cell part, organelle, and membrane components, indicating that a large proportion of the transcripts encode structural and compartmental proteins. Approximately 13% of the unigenes had a hit to proteins included in at least one database. A smaller subset was linked to membrane-enclosed lumen, extracellular region, and protein-containing complexes, suggesting involvement in intercellular communication and specialized organellar functions. The Biological Process category encompassed a wide array of pathways, with enrichment in metabolic processes, cellular processes, biological regulation, response to stimulus, and developmental processes. The dominance of genes under metabolic and biosynthetic pathways suggests active transcriptional regulation

supporting cellular growth, stress response, and secondary metabolite production.

The GO annotation shows that the butterfly pea transcriptome is functionally diverse, with strong representation in the Core cellular structures (CC), Catalysis, binding, and regulatory roles (MF) and Development, metabolism, and pigmentation-related processes (BP). Further, the GO classification indicates that most genes in the butterfly pea transcriptome are involved in essential cellular and metabolic functions, with significant subsets associated with transcriptional regulation, pigment biosynthesis, and stress responses (Table 1, Data File 9 [27]). These functional annotations provide a basis for linking differentially expressed genes to pathways regulating flower color diversity among varieties.

Limitations

The transcriptome assembly was obtained from RNA extracted from three different biological replicates by ensuring consistent time of harvest and fresh first-day blooms. However, environmental factors or plant physiological state as well as potential plant stress due viral infection may contribute to the expression profile.

Abbreviations

NCBI	National Centre for Biotechnology
SRA	Sequence Read Archive
GO	Gene Ontology
EC	Enzyme Codes
PB	Blue flower variant of butterfly pea
PW	White flower variant of butterfly pea
SB	Sky or light blue flower variant of butterfly pea

Acknowledgements

This research was financially supported by Taylor's University—Taylor's Impact Lab Grant—TIRGS-ILG/1/2023/SOB/003.

Author contributions

YWH and AB conceived and designed the study; YWH funding acquisition; NS performed sample collection, RNA extraction, and library preparation; MAMA conducted bioinformatic analysis and data visualization; MAMA contributed to data interpretation and figure preparation; NS drafted the manuscript; MAMA and YWH critically revised and finalized the manuscript; All authors read and approved the final version of the manuscript.

Funding

This study was funded by Taylor's University—Taylor's Impact Lab Grant—TIRGS-ILG/1/2023/SOB/003.

Data availability

The data described in this Data Note can be freely and openly accessed in the NCBI and fig share databases. The sequencing data (with accession numbers SRX30672033– SRX30672041 under BioProject PRJNA1334233) have been deposited in the SRA database of NCBI under SRP62704, with the Transcriptome Shotgun Assembly accession number GLLN00000000. The other data files generated in the current study are available in the Figshare database. Please see Table 1 and references for details and links to the data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 30 October 2025 / Accepted: 23 April 2026
Published online: 30 April 2026

References

1. Jing P, Giusti MM. Effects of extraction conditions on improving the yield and quality of an anthocyanin-rich purple corn (*Zea mays* L.) color extract. J Food Sci. 2007;72(6):C363–8. https://doi.org/10.1111/j.1750-3841.2007.00409.x.
2. Salehi B, Sharifi-Rad J, Cappellini F, Reiner Ž, Zorzan D, Imran M, et al. The therapeutic potential of anthocyanins: current approaches based on their molecular mechanism of action. Front Pharmacol. 2020;11:1300. https://doi.org/10.3389/fphar.2020.01300.
3. Khoo HE, Azlan A, Tang ST, Lim SM. Anthocyanidins and anthocyanins: colored pigments as food, pharmaceutical ingredients, and the potential health benefits. Food Nutr Res. 2017;61(1):1361779. https://doi.org/10.1080/16546628.2017.1361779.
4. Gonçalves AC, Romano A, Martins I, Reis F, Bronze MR. Dietary effects of anthocyanins in human health. Antioxidants. 2021;10(7):1117. https://doi.org/10.3390/antiox10071117.
5. Yoon J, Hogue A, Shearer G. Anthocyanins as a promising phytochemical and dietary therapeutic agent for the treatment of colorectal cancer. Nutrients. 2018;10(6):849. https://doi.org/10.3390/nu10060849.
6. Li J, Li S, Li B, Wang Y. Anthocyanins and their biosynthesis in plants. Plant Physiol J. 2019;55(3):260–71.
7. Yue J, Yang Y, Li L, Henry RJ. Anthocyanins in plants: Biosynthesis, regulation and health benefits. Molecules. 2019;24(13):2593. https://doi.org/10.3390/molecules24132593.
8. Kazuma K, Noda N, Suzuki M. Flavonoid composition related to petal color in different lines of *Clitoria ternatea*. Phytochemistry. 2003;64(6):1285–90.
9. Gamage GCV, Lim YY, Choo WS. Anthocyanins from *Clitoria ternatea* flower: Biosynthesis, extraction, stability, antioxidant activity, and applications. Front Plant Sci. 2021;12:792303. https://doi.org/10.3389/fpls.2021.792303.
10. Abidin NAZ, Aziz A, Alias Y, Isa AH, Zamri MFM. Effects of sunflower pectin on purple sweet potato anthocyanins during different pH. Food Chem. 2019;312:126074. https://doi.org/10.1016/j.foodchem.2019.126074.
11. Thuy NMT, Minh VQ, Ben TC, Thi Nguyen MT, Ha HTN, Tai NV. Identification of anthocyanin compounds in butterfly pea flowers (*Clitoria ternatea* L.) by ultra performance liquid chromatography/ultraviolet coupled to mass spectrometry. Molecules. 2021;26(15):4539. https://doi.org/10.3390/molecules26154539.
12. Yoshida K, Mori M, Kondo T. Blue flower color development by anthocyanins: from chemical structure to cell physiology. Nat Prod Rep. 2009;26(7):884–915. https://doi.org/10.1039/b800165k.
13. Zhao L. Anthocyanin pigment biosynthesis, regulation, and applications in plants: a review. Plant Physiol Biochem. 2021;160:215–27. https://doi.org/10.1016/j.plaphy.2021.06.022.
14. Liu Y, Tikunov Y, Schouten RE, Marcelis LFM, Visser RGF, Bovy A. Anthocyanin biosynthesis and degradation mechanisms in Solanaceous vegetables: a review. Front Chem. 2018;6:52. https://doi.org/10.3389/fchem.2018.00052.
15. Fareed FS, Singaram N. A modified RNA extraction protocol for secondary metabolite-rich *Amomum* species (Zingiberaceae). Malays J Biochem Mol Biol. 2023;26(2):51–8.
16. Singaram N. Extracted RNA quantity, quality and RIN value. figshare. 2025. https://doi.org/10.6084/m9.figshare.30487016.v5.
17. Yin S, Cheng Y, Liu H, Zhang R, Wang J. An efficient and effective RNA extraction protocol for ferns. Appl Plant Sci. 2024;12(9):e11604. https://doi.org/10.1002/aps3.11604.
18. Walker RM, Brown RH, Rattray M, Sidders B, Fox NC, Lewis PA. Quality control, modeling, and visualization of RNA-Seq data using the RIN (RNA Integrity Number) metric. Biotechniques. 2016;60(1):13–20. https://doi.org/10.2144/000114374.
19. Singaram N, Mohamed Akbar MAB, Bunawan H, Yap WH. RNA-seq data from white, light-blue, and dark-blue *Clitoria ternatea* flowers. NCBI Sequence Read Archive. 2025. https://identifiers.org/ncbi/insdc.sra:SRP627045. Accessed 26 Nov 2025.

20. Singaram N. Transcriptome de novo assembly and annotation methodology. figshare. 2025. https://doi.org/10.6084/m9.figshare.30486929.v2.
21. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 26 Dec 2025.
22. Singaram N. MultiQC read quality results. figshare. Figure. 2025. https://doi.org/10.6084/m9.figshare.30472619.v4.
23. Singaram N. RNA-seq De novo transcriptome assembly. figshare. 2025. https://doi.org/10.6084/m9.figshare.30472562.v2.
24. Singaram N. TSA: *Clitoria ternatea* cultivar Butterfly pea isolate, transcriptome shotgun assembly. 2025. https://identifiers.org/ncbi/insdc:GLLN00000000.1. Accessed 2 Jan 2026.
25. Singaram N. Protein sequence prediction. figshare. 2025. https://doi.org/10.6084/m9.figshare.30472592.v2.
26. Singaram N. Gene Ontology annotation. figshare. Figure. 2025. https://doi.org/10.6084/m9.figshare.30465206.v3.
27. Singaram N. Functional annotation. figshare. 2025. https://doi.org/10.6084/m9.figshare.30486899.v2.

Publisher's Note