Cannabis Expression Atlas: a comprehensive resource for integrative analysis of *Cannabis sativa* L. gene expression

Kevelin Barbosa-Xavier[1], Francisnei Pedrosa-Silva[1], Fabricio Almeida-Silva[2,3], Thiago M. Venancio[1]*

[1] Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.

[3] VIB Center for Plant Systems Biology, VIB, Ghent, Belgium.

* TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com.

Abstract

Cannabis sativa L., a plant originating from Central Asia, is a versatile crop with applications spanning textiles, construction, pharmaceuticals, and food products. This study aimed to compile and analyze publicly available Cannabis RNA-Seq data and develop an integrated database tool to help advance Cannabis research in various topics such as fiber production, cannabinoid biosynthesis, sex determination, and plant development. We identified 515 publicly available RNA-Seq samples that, after stringent quality control, resulted in a high-quality dataset of 394 samples. Utilizing the Jamaican Lion genome as reference, we constructed a comprehensive database and developed the Cannabis Expression Atlas (https://cannatlas.venanciogroup.uenf.br/), a web application for visualization of gene expression, annotation, and functional classification. Key findings include the quantification of 27,640 Cannabis genes and their classification into seven expression categories: not-expressed, low-expressed, housekeeping, tissue-specific, group-enriched, mixed, and expressed-in-all tissues. The study revealed substantial variability and coherence in gene expression across different tissues and chemotypes. We found 2,382 tissue-specific genes, including 177 transcription factors. . The Cannabis Expression Atlas constitutes a valuable tool for exploring gene expression patterns and offers insights into Cannabis biology, supporting research in plant breeding, genetic engineering, biochemistry, and functional genomics.

Keywords: transcriptomics, bioinformatics, integrative biology, analysis tool.

1-Introduction

Cannabis sativa L., commonly referred to as Cannabis, hemp, marijuana, or pot, originates from Central Asia and stands among the most versatile crops, serving as a fundamental resource for textiles, paper, construction materials, pharmaceuticals, and food products (Gao *et al.*, 2020). With a cultivation history spanning over 10,000 years, Cannabis has been integral to textile production in China for over 6,000 years (Hussain *et al.*, 2021). Evidence from Central Asia indicates that the utilization of drug-type genotypes, particularly for medicinal and ritualistic purposes, dates back to over 2,700 years (Ren *et al.*, 2019).

Currently, the global market for hemp products generates over 100 million US dollars (US$) annually (United Nations, 2024). From 2019 to 2022, the global export of hemp seeds averaged 122 million US$, with Canada being the largest exporter, accounting for 46.54% in 2022. By 2022, 41 countries were exporting hemp seeds or related products (United Nations, 2024). Meanwhile, hemp fibers averaged 14.4 million US$ in trade between 2019 and 2022. During the same period, an average of 40 countries imported hemp fiber products, and about 25 countries exported them, with European countries dominating the market (United Nations, 2024).

Cannabis has a diploid genome (2n = 20) comprising nine pairs of autosomes and one pair of sex chromosomes, rendering it a dioecious species wherein male and female plants exhibit XY and XX sex chromosomes, respectively (Grassa *et al.*, 2018). The Cannabis Y chromosome has an entire heterochromatic arm and a large pseudoautosomal region on the other arm indicating that some genes have homologs at the X chromosome and have autosomal behavior (Divashuk *et al.*, 2014).

The primary feature that makes Cannabis widely utilized for medicinal, ceremonial, and adult purposes is its diverse cannabinoid profile (ElSohly and Slade, 2005). Several cannabinoid compounds have been identified, with Tetrahydrocannabinol (THC) and Cannabidiol (CBD) emerging as the most widely recognized. Based on their cannabinoid profiles, Cannabis plants can be categorized into different chemotypes (de Meijer *et al.*, 2003). Type I, commonly known as marijuana, primarily contains tetrahydrocannabinolic acid (THCA) as the predominant compound. Type II plants exhibit similar levels of THCA and cannabidiolic acid (CBDA). Finally, type III plants, also known as hemp, feature CBDA as the primary compound. Cannabis plants synthesize THCA and CBDA, which are decarboxylated to THC and CBD, respectively. Legislation in various countries often relies on THC concentration to distinguish between marijuana and hemp. Typically, plants containing more than 0.3% THC are classified as marijuana, while those with less than 0.3% THC are considered hemp (Stout *et al.*, 2012).

Scientific progress in Cannabis research, including DNA and RNA sequencing, has greatly improved our comprehension of the molecular mechanisms underlying fiber production, cannabinoid biosynthesis, sex determination, and plant development (Guerriero *et al.*, 2017; McKernan *et al.*, 2020; Adal *et al.*, 2021; Tang *et al.*, 2023). Over the past 13 years, several research groups have generated a substantial volume of Cannabis RNA-Seq data (Bakel *et al.*, 2011; Guerriero *et al.*, 2017; Prentout *et al.*, 2020). Nevertheless, there is a lack of systematic integrative analyses of these datasets, which often require specialized personnel and computational resources beyond the means of most research groups. Moreover, databases with visualization tools that integrate gene expression data from multiple chemotypes and tissues are crucial to accelerate research projects, and empower Cannabis researchers worldwide.

In this study, we performed an extensive analysis of all publicly available Cannabis RNA-Seq data, resulting in a curated dataset of 394 high-quality samples. We estimated transcriptional abundances of 27,640 genes, including 1,465 transcription factors (TFs). We further classified these genes into seven expression categories, including tissue-specific and housekeeping, providing valuable insights into Cannabis functional genomics. To enhance accessibility and usability, we developed a user-friendly web application, the Cannabis Expression Atlas (Cannatlas; https://cannatlas.venanciogroup.uenf.br/), that enables researchers to explore gene expression and molecular biology of Cannabis, supporting plant breeding, genetic engineering, and medicinal research.

2-Materials and Methods

Data acquisition, quality check, and transcript abundance estimation

We identified the available Cannabis RNA-Seq data on the Sequence Read Archive (SRA) database using the search parameters “Cannabis sativa”[organism] AND “rna-seq”[strategy]. We used the R package ‘bears’ (Almeida-Silva, Pedrosa-Silva and Venancio, 2023) to obtain sample metadata with the create_sample_info() function. The FASTQ files were downloaded from the European Nucleotide Archive’s mirror of SRA using the download_from_ena() function, and file integrity was verified with the check_md5() function. Adapters and low quality bases were removed using FASTP (Chen *et al.*, 2018). At this stage, we excluded samples with mean read length of less than 40 or a Q20 rate below 80% after filtering (Almeida-Silva, Pedrosa-Silva and Venancio, 2023).

The most commonly used Cannabis reference genomes are Finola, Purple Kush, and Cs10 (Grassa *et al.*, 2018; Laverty *et al.*, 2019), which belong to type I and III female plants. This has two implications: i. they do not have male-specific genome segments, and ii. type I plants do not have a functional CBDAS gene, while type III plants do not have a functional THCAS gene (McKernan *et al.*, 2020). McKernan *et al.* (2020) reported the genome of a type II female and male plant, the Jamaican Lion Mother (JL mother) and the Jamaican Lion Father (JL father), and found male-specific contigs, making this the most complete Cannabis genome to date. The JL mother and JL father are available in GenBank under accession numbers GCA_012923435.1 and GCA_013030025.1, respectively. We used as reference the JL mother genome, complemented with the Y contigs from the JL father (McKernan *et al.*, 2020) as a reference in our atlas. This genome was used to build a reference transcriptome with the R packages GenomicFeatures and Rsamtools (Lawrence *et al.*, 2013; Martin Morgan *et al.*, 2023).

We estimated gene-level transcript abundances with ‘salmon’ (Patro *et al.*, 2017), which performs pseudo-mapping to the reference transcriptome followed by quantification. At this stage, we excluded samples with less than 50% mapped reads (Almeida-Silva, Pedrosa-Silva and Venancio, 2023). Additionally, genes with less than 1 transcript per million (TPM) in all samples were classified as not-expressed and not used in the downstream analysis.

Functional gene annotation

Functional gene annotation was conducted using an integration of three database tools: InterProScan 5, UniProt - IDmapping, and KEGG - GhostKOALA (Jones *et al.*, 2014; Kanehisa, Sato and Morishima, 2016; The UniProt Consortium, 2023). These tools enabled us to construct an annotation matrix with the following information: Gene ID, Protein ID, Entry name (UniProt), Protein name (UniProt), Gene Ontology (UniProt), Accession (InterPro), Description (InterPro), Gene Ontology (InterPro), Pathway annotations (InterPro), Entry (KEGG), Definition (KEGG), Pathway (KEGG), and Module (KEGG). TF prediction was conducted using PlantTFDB v5.0 (Tian *et al.*, 2020).

Dimensionality reduction

Dimensionality reduction methods were used to evaluate sample clustering patterns. The ‘scran’ package (Lun, McCarthy and Marioni, 2016) was used to model the mean-variance relationship in a matrix of log-transformed counts normalized by library size, followed by the identification of the top 5000 genes with the highest expression variability. Next, we performed a principal component analysis (PCA) and used the elbow point statistic to extract the top principal components. We employed the t-stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton, 2008) dimensionality reduction method to compare the data distribution in a low-dimensional space. We tested 6 different perplexity values (10, 20, 30, 40, 50, and 60) and selected 30 as the optimal value based on visual inspection.

Gene classification by expression category

We identified tissue-specific genes for tissues with at least 10 samples (leaf, trichome, female flowers, roots, hypocotyls, stem, male flowers, seeds, and bast fibres). Although induced male flowers had only 6 samples, we kept them to identify tissue-specific genes. Samples from mixed tissues were excluded. We obtained the median TPM expression of each gene in each tissue, as previously described by (Machado *et al.*, 2020; Almeida-Silva, Pedrosa-Silva and Venancio, 2023). Genes with median TPM of at least 5 in a tissue were classified using the ‘TissueEnrich’ R package (Jain and Tuteja, 2019) and had their tissue specificity index TAU calculated using the following formula:

$$\tau = \frac{\sum_{i=1}^{n} (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max\limits_{1 \le i \le n} (x_i)}$$

where:
x_i = gene expression in tissue i.
n = number of tissues.

Genes identified as “Tissue-Enriched” or “Tissue-Enhanced” by the teGeneRetrival() function of the ‘TissueEnrich’ package and with TAU ≥ 0.8 were classified as “tissue-specific” genes.

Furthermore, 'TissueEnrich' also returns other three gene expression categories, the "Group-Enriched", "Mixed", and "Expressed-in-all".

Identification of housekeeping genes

To consistently identify housekeeping genes, we utilized the methods previously employed in the Soybean Expression Atlas developed by our group (Machado *et al.*, 2020; Almeida-Silva, Pedrosa-Silva and Venancio, 2023):

- Selected the genes with TPM ≥ 5 in at least one sample;
- Selected genes that are expressed in all samples;
- Obtained the median TPM of each gene across all samples;
- Computed the standard deviation (sd) and the coefficient of variation (CoV) of gene expression;
- Computed the maximum fold change (MFC) by determining the largest and smallest TPM value;
- The MFC-CoV score was calculated by multiplying the MFC with the CoV;
- Identified the first quartile;
- Identified the Tau index for the housekeeping putative genes.

Genes with MFC-CoV score within the first quartile and with a Tau index ≤ 0.4 were classified as housekeeping genes.

Data visualization

Data visualization was mainly performed using ggplot2 (Wickham, 2016), Pheatmap (Kolde, 2019) and plotly (Plotly Technologies Inc, 2015).

Web application development

We built a web application using 'Shiny' with the shinydashboard template (Chang *et al.*, 2017; Chang and Ribeiro, 2018). The gene expression database used in the application is stored in a partitioned parquet directory. The interface between R and the Apache Arrow platform is performed with the 'Arrow' R package (Richardson *et al.*, 2021). The app is freely accessible at https://cannatlas.venanciogroup.uenf.br/.

3-Resource overview

We developed the Cannabis Expression Atlas, a web application designed to explore the expression patterns of Cannabis genes using RNA-Seq data. This atlas currently comprises the expression of 27,640 genes across 394 RNA-seq samples, along with sample metadata (Table S1). It consists of four main pages, each offering different data navigation options (Figure 1):

I. **Gene explorer:** users can search individual Gene IDs and explore their expression profiles through t-SNE, median, and mean TPM barplots. A representative image of the tissues colored according to the median expression levels of the gene of interest is also provided.

II. **Batch search:** users can fetch the expression of multiple genes. In this case, there are two navigation options:

   A. **Gene information tab:** users can submit gene lists and explore a dynamic table that can be filtered by Gene ID, classification, tissue, GO terms, pathways, among other annotation information. Different plots, statistics, and download options are also provided.

   B. **Gene expression tab:** users can submit a list of genes and explore their expression across tissues, represented in a heatmap that is generated along an accessory table with sample metadata.

III. **BLAST Search:** users can find genes of interest by querying nucleotide or amino acid sequences.

IV. **Batch download:** in this page, users can fetch expression data by one of two tabs:

   A. **Download by tissue:** selection of the tissue of interest and quantification method (TPM or counts);

   B. **Download by BioProject:** users can filter a table by cultivar, chemotype, or publication keywords, DOI, PMID, and title. The web app then generates an expression matrix and a file with the relevant metadata.

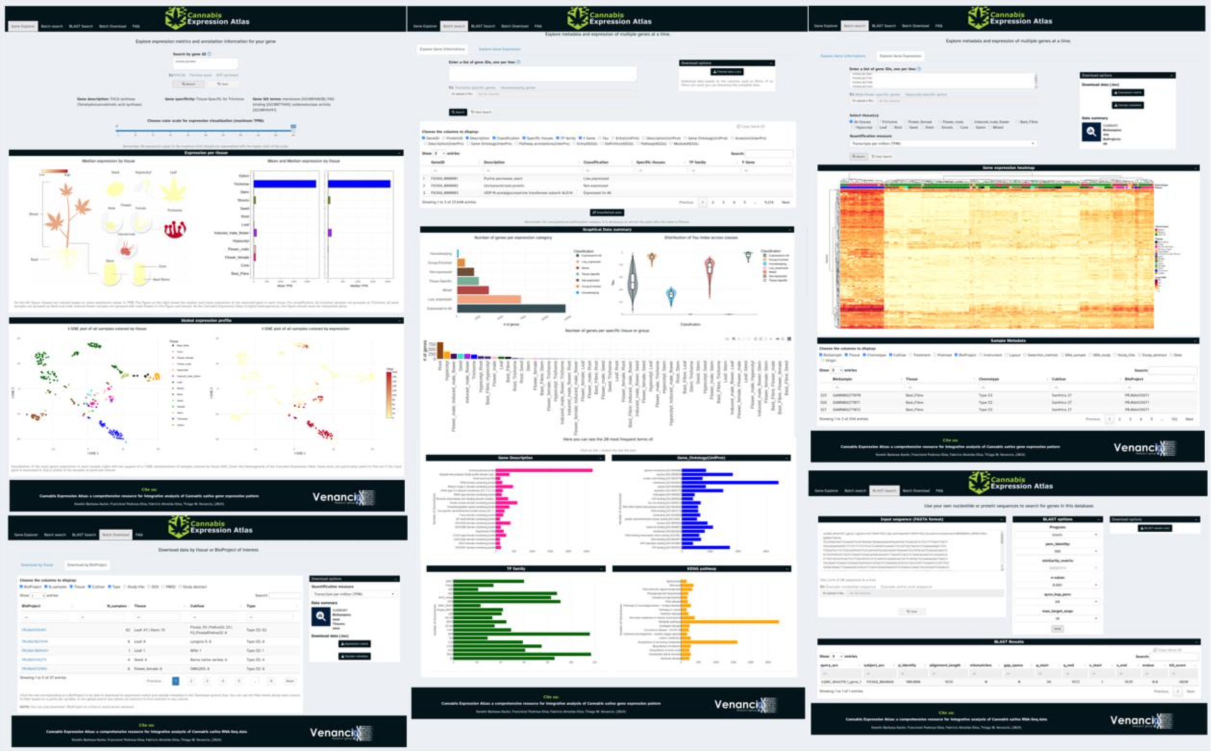

Figure 1. Overview of the Cannabis Expression Atlas.

4-Results and discussion

Publicly available Cannabis RNA-seq data comprise 394 high-quality samples of 13 tissues, 55 varieties, and three chemotypes

In August 2024, we identified 515 publicly available RNA-Seq samples from the SRA database. Following download and preprocessing, which included read trimming to remove adapters and low-quality bases, we excluded 40 samples that exhibited a mean length of less than 40 bp or a Q20 rate below 80% (Figure S1a, b, c). These samples belong to only four bioprojects (Table S4a). We also removed 81 samples with read mapping rates lower than 50% (Figure S1a, d; Table S4b), resulting in a database of 394 high-quality samples. By retaining only high-quality samples, we improve the reliability of gene expression quantification and minimize the prevalence of potentially contaminated samples (Almeida-Silva, Pedrosa-Silva and Venancio, 2023). By using this filtered dataset, we successfully quantified transcript abundance of 27,640 Cannabis genes representing 99% of the reference transcriptome (JL (27,358) + Y genes (535) = 27,893 genes).

The atlas encompasses 13 different plant tissues and the mixed samples (Figure 2a, c), with leaf representing 38.3% of the samples. We grouped these 394 samples into the three main chemotypes (Figure 2b), namely type I (marijuana), type II (balanced), and type III (hemp). Furthermore, these samples originate from 55 Cannabis varieties with the four most represented being Finola (n = 42), Santhica 27 (n = 36), a F1 from Agp029 x blackberry kush auto (n = 27), and Felina32 (n = 23) (Figure 2d), of which the F1 from Agp029 x blackberry kush is a type I and the other three are type III. The summary statistics reveal that 76.5% of the Cannabis RNA-seq data available in public databases are of good quality, with the mapping rate filter being responsible for the exclusion of 15.7% of the samples (Figure S1a, d). Most samples demonstrated high mapping rates, with an average of approximately 80% and a smaller cluster around 40% (Figure S1d). The median and mean number of reads per sample were calculated at 42 and 51 million reads, respectively (Figure S1b). Paired-end sequencing was used for most samples (366), while the Illumina NovaSeq 6000 platform was employed for 153 samples.

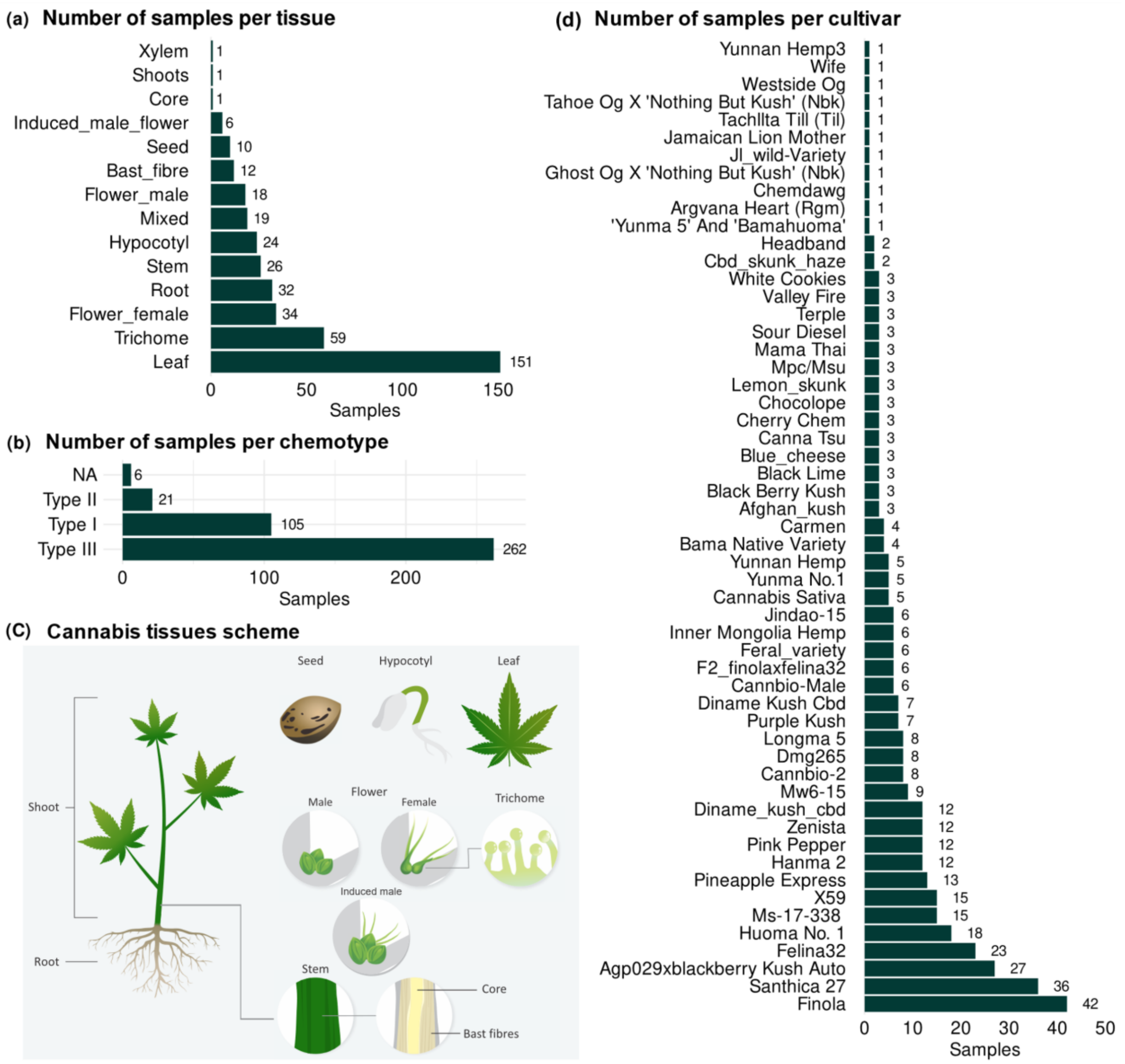

Figure 2. Distribution of Cannabis samples. (a) Number of samples per tissue. (b) Number of samples per chemotype. (c) Schematic representation of Cannabis tissues. (d) Number of samples per cultivar.

The increase in RNA-Seq data highlights the growing interest in Cannabis research

The first publication of a Cannabis RNA-seq study dates back to 2011 (Bakel *et al.*, 2011). However, a significant number of studies emerged only after 2016 (Figure S1e, f, g), coinciding with the global legislative shift on Cannabis use and research (United Nations, 2024). Notably, type III (hemp) samples outnumbered type I (marijuana) and type II (balanced) samples (Figure S1f), reflecting industrial and agricultural interests. Interestingly, since 2019, there has been a notable increase in the number of samples derived from plant tissues associated with medicinal applications, such as trichomes and female flowers, which correlates with an increase in types I and II samples (Figure S1f, g) which highlights the growing emphasis on the Cannabis therapeutic potential. Over 51% of the samples were contributed by research groups based in China and Canada (Figure S1h). These RNA-Seq studies addressed a myriad of Cannabis research questions, including fiber production and quality (Guerriero *et* *al.*, 2017), biotic and abiotic resistance (Gao *et al.*, 2018; McKernan *et al.*, 2020; Cao *et al.*, 2021, 2023; Jiang *et al.*, 2021; Pépin, Hebert and Joly, 2021; Yin *et al.*, 2022; Yan *et al.*, 2023), sex determination (Prentout *et al.*, 2020; Adal *et al.*, 2021; Dowling *et al.*, 2023), metabolite production, quality, chemotype identification (Braich *et al.*, 2019; Laverty *et al.*, 2019; Zager *et al.*, 2019; Booth *et al.*, 2020; Livingston *et al.*, 2020; McGarvey *et al.*, 2020; McKernan *et al.*, 2020; Busta *et al.*, 2022; Yeo *et al.*, 2022; Mi *et al.*, 2023; Tang *et al.*, 2023), and genome assembly (Bakel *et al.*, 2011; Braich *et al.*, 2020; Gao *et al.*, 2020; McKernan *et al.*, 2020).

Dimensionality reduction reveals distinct transcriptional profiles by tissue and chemotype

The elbow point statistics indicated that 8 principal components account for 65% of the expression variation. The t-SNE dimensionality reduction using these components shows that the most significant variation associated with chemotype was observed in leaves, where two distinct clusters of type I samples were separated from type III samples (Figure 3). The two Type I leaf clusters are of two different varieties and BioProjects, with one (the bigger leaf type I cluster, Figure 3) being focused on resistance to *Golovinomyces ambrosiae* (PRJNA738505) and the other (the smallest leaf type I cluster, Figure 3) focused on study auto-flowering (PRJNA1049889). Fiber-related tissues, such as hypocotyl, bast fiber, and stem formed a tightly grouped cluster, reflecting their shared role in structural support and potential similarity in gene expression related to cell wall biosynthesis and secondary growth. Reproductive tissue samples (female, male and induced male flowers) clustered closely although not very cohesively (Figure 3) and a divergence between female and male flowers is noticeable. Chemotype and maturity associated clustering patterns were also observed. Trichomes exhibited comparable global expression patterns regardless of chemotype. Finally, root samples were clearly distinct from all other tissue types.

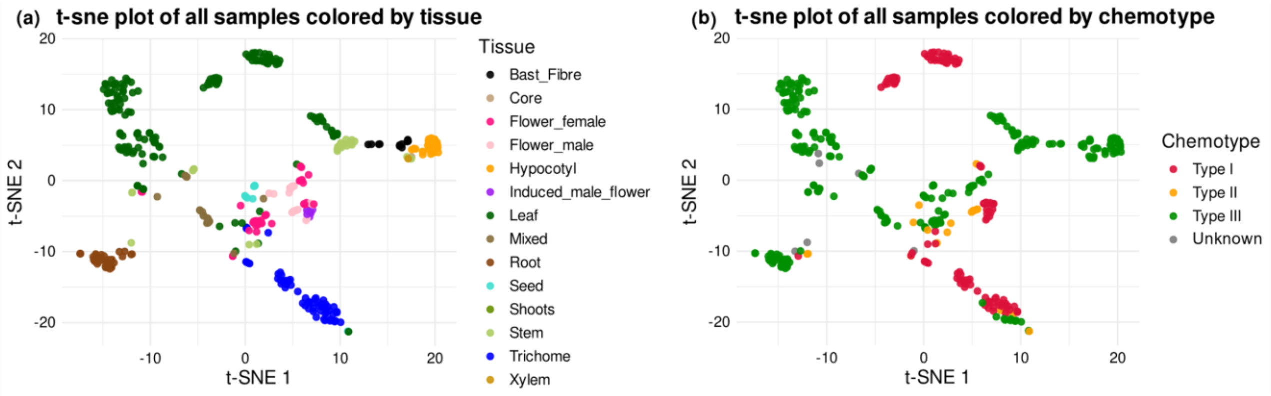

Figure 3. Dimensionality reduction performed with t-SNE. (a) Samples colored by tissue. (b) Samples colored by chemotype.

The Cannabis Expression Atlas comprises seven gene expression categories and over 1400 transcription factors

We classified genes into seven expression categories (Figure 4, Table S2) (Jain and Tuteja, 2019; Machado *et al.*, 2020; Almeida-Silva, Pedrosa-Silva and Venancio, 2023). Housekeeping genes (HK, 132 genes) were identified as those with consistently high expression and low variation across samples (Figure S2a) which support their roles in fundamental cellular functions and physiological processes (Joshi *et al.*, 2022). Tissue-specific genes (2,382 genes) were defined by a TAU index between 0.8 and 1.0 and at least five-fold higher expression in one tissue compared to others (Figure 4 and Figure S2b). These genes can shed light on the specific regulatory mechanisms governing the development of various plant parts. Group-enriched genes (843 genes) have at least five-fold higher expression levels in a group of 2 tissues in comparison to all other tissues. The Expressed-in-all category comprises 11,929 genes expressed in all tissues. The Low-expressed group (7,021 genes) had TPM $\geq$ 1 in at least one sample but median TPM $\leq$ 5, while Not-expressed genes (1,853) had TPM < 1 in all samples. Finally, 3,480 genes were classified as mixed, not fitting into the previous categories.

By utilizing the reference genome (see methods for details), we identified 535 Y-linked genes of the JL father, termed Y-genes for brevity. Of these, 370 had detectable expression in our dataset (see methods for details), out of which 130 showed median TPM $\geq$ 5 and were analyzed for tissue-specific expression. From Y-genes, 240 were Low-expressed, 165 were Not-expressed, 61 were specifically expressed in male flowers, 26 showed mixed expression, 22 were expressed-in-all and 11 were group-enriched (Table S2). The Y-genes classified as mixed or expressed-in-all suggest that they are located in the pseudoautosomal region of Y chromosome (Divashuk *et al.*, 2014). This list of Y-genes and their expression patterns can offer insights into sex determination and male flower development and physiology.

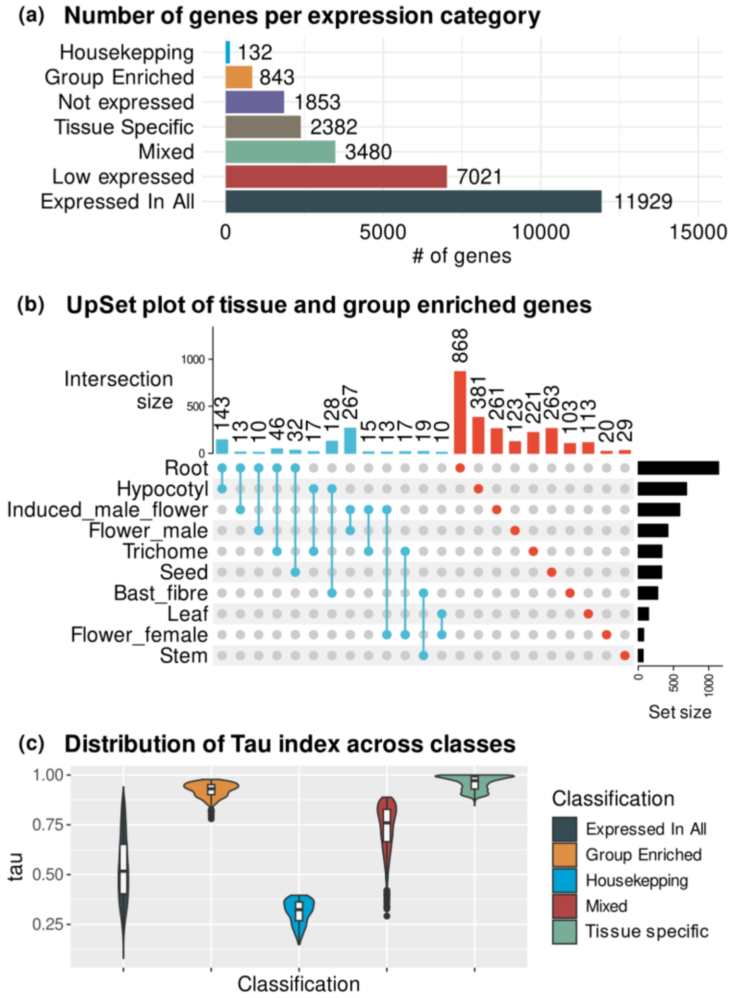

Figure 4. Gene expression classification. (a) Number of genes per expression category. (b) UpSet plot of tissue and group specific genes. (c) Violin plot of distribution of the Tau index across gene classes.

Out of the 27,640 genes in the Cannabis Expression Atlas, we found 1,489 TFs using PlantTFDB (Table S2) as a reference, out of which 177 (11.8%) were tissue-specific (Table 1). The top 10 TF families in number of genes were bHLH, ERF, MYB, NAC, B3, MYB_related, C2H2, bZIP, WRKY, and C3H (Table S3) and correspond to 56.1% of all TFs of the Cannabis Expression Atlas. The highest prevalence of tissue specific TFs was observed in roots (n = 75), followed by hypocotyl (n = 30), seed (n = 21), and induced male flowers (n = 17) (Table 1). This pattern aligns with the high number of specific genes identified in these tissues (Figure 4b). These TFs can be important regulators of tissue-specific regulatory mechanisms and constitute important candidates for biotechnological applications.

Table 1. Tissue-specific transcription factors.

TF family	Root	Hypocotyl	Seed	Induced male flower	Trichome	Bast fibre	Flower female	Flower male	Stem	Leaf	Total
MYB	8	5	1	3	2	2	2				**23**
ERF	14	1	4	1	2						**22**
WRKY	15				1				2		**18**
bHLH	9	2		3		1		2			**17**
NAC	4	9		2	1						**16**
LBD	5	3	1	3					1		**13**
MYB_related	4	1		1	2						**8**
MIKC_MADS	1				1	1	3	1			**7**
C2H2	3	3									**6**
bZIP	2		1			1		1			**5**
AP2		1	2		1						**4**
B3			3	1							**4**
G2-like	2	1	1								**4**
WOX	1	1	2								**4**
GRAS	3										**3**
M-type_MADS	1		1			1					**3**
SRS					3						**3**
C3H		1	1								**2**
HSF	1		1								**2**
NF-YB	1		1								**2**
SBP				1		1					**2**
YABBY				2							**2**
CO-like										1	**1**
Dof	1										**1**
GRF			1								**1**
GeBP		1									**1**
S1Fa-like			1								**1**
TALE		1									**1**
Trihelix								1			**1**
Total	**75**	**30**	**21**	**17**	**13**	**7**	**5**	**5**	**3**	**1**	**177**

Interestingly, the Cannabis Expression Atlas has more TFs than that reported in PlantTFDB 5.0 for Cannabis (Tian *et al.*, 2020). Table 2 presents the top 10 TF families in the Cannabis Expression Atlas. This discrepancy may be attributed to the fact that the Cannabis TFs in PlantTFDB 5.0 are based on the Purple Kush genome, which has just 78.1% of BUSCO completeness score, whereas the one used here achieved 96.8% (Medicinal Genomics, 2019). The progress in genome sequencing and assembly likely impacted gene predictions.

Table 2. Top 10 transcription factors family, in number of genes, at Cannabis Expression Atlas compared to PlanTFDB.

TF_family	Cannabis Expression Atlas members	PlantTFDB Cannabis Members (v5.0)	Fold change (Cannatlas/PlantTFDB)
bHLH	117	99	1.18
ERF	116	59	1.96
MYB	91	81	1.12
NAC	88	75	1.17
B3	86	60	1.43
MYB_related	83	70	1.18
C2H2	75	62	1.20
bZIP	61	54	1.12
WRKY	60	49	1.22
C3H	59	48	1.22
Total	**836**	**657**	**1.27**

5-Conclusions

The development of the Cannabis Expression Atlas represents significant progress in the field, providing researchers with a user-friendly and freely available platform to explore gene expression data. This tool facilitates the investigation of gene expression patterns across different tissues and conditions, enabling researchers to conduct in-depth analyses and generate new hypotheses. The integration of metadata and various search options enhances the usability of the atlas, making it a valuable resource for the research community.

Author contributions

Conceived the study: K.B.-X. and T.M.V.; Funding and resources: T.M.V.; Data analysis: K.B.-X.; Interpretation of the results: K.B.-X. and T.M.V.; Source development: K.B.-X., F.A.-S., F.P.-S. and T.M.V.; Wrote the manuscript: K.B.-X. and T.M.V.

Acknowledgment

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq). The funding agencies had no role in the design of the study and collection, analysis and interpretation of data and in writing.

Data availability and FAIR (Findable Accessible Interoperable Reusable) compliance statement

The Cannabis Expression Atlas container is available at Docker hub and the parquet directory is available at FigShare. All the data files, supplementary materials and methodology code used here are available at GitHub.

Reference list

Adal, A.M. *et al.* (2021) 'Comparative RNA-Seq analysis reveals genes associated with masculinization in female Cannabis sativa', *Planta*, 253(1). Available at: https://doi.org/10.1007/S00425-020-03522-Y.

Almeida-Silva, F., Pedrosa-Silva, F. and Venancio, T.M. (2023) 'The Soybean Expression Atlas v2: A comprehensive database of over 5000 RNA-seq samples', *The Plant Journal*, 116(4), pp. 1041–1051. Available at: https://doi.org/10.1111/tpj.16459.

Bakel, H. van *et al.* (2011) 'The draft genome and transcriptome of Cannabis sativa.', *Genome biology*, 12(10), p. R102. Available at: https://doi.org/10.1186/gb-2011-12-10-r102.

Booth, J.K. *et al.* (2020) 'Terpene Synthases and Terpene Variation in Cannabis sativa1[OPEN]', *Plant Physiology*, 184(1), pp. 130–147. Available at: https://doi.org/10.1104/pp.20.00593.

Braich, S. *et al.* (2019) 'Generation of a Comprehensive Transcriptome Atlas and Transcriptome Dynamics in Medicinal Cannabis.', *Scientific reports*, 9(1), p. 16583. Available at: https://doi.org/10.1038/s41598-019-53023-6.

Braich, S. *et al.* (2020) 'A new and improved genome sequence of Cannabis sativa', *Gigabyte*, 2020, pp. 1–13. Available at: https://doi.org/10.46471/gigabyte.10.

Busta, L. *et al.* (2022) 'Chemical and genetic variation in feral *Cannabis sativa* populations across the

Nebraska climate gradient', *Phytochemistry*, 200, p. 113206. Available at: https://doi.org/10.1016/j.phytochem.2022.113206.

Cao, K. *et al.* (2021) 'The transcriptome of saline-alkaline resistant industrial hemp (Cannabis sativa L.) exposed to NaHCO3 stress', *Industrial Crops and Products*, 170, p. 113766. Available at: https://doi.org/10.1016/J.INDCROP.2021.113766.

Cao, K. *et al.* (2023) 'The miRNA–mRNA regulatory networks of the response to NaHCO3 stress in industrial hemp (Cannabis sativa L.)', *BMC Plant Biology*, 23, p. 509. Available at: https://doi.org/10.1186/s12870-023-04463-w.

Chang, W. *et al.* (2017) 'Shiny: web application framework for R'. Available at: https://scholar.google.com/scholar?q=Chang%2C+W.+and+Ribeiro%2C+B.B.+%282019%29+shinydashboard%3A+Create+Dashboards+with+"Shiny".+R+package+version+0.7.1. (Accessed: 3 December 2023).

Chang, W. and Ribeiro, B.B. (2018) 'shinydashboard: Create Dashboards with "Shiny"'. Available at: https://scholar.google.com/scholar?q=Chang%2C+W.%2C+Cheng%2C+J.%2C+Allaire%2C+J.%2C+et+al.+%282021%29+shiny%3A+Web+Application+Framework+for+R.+2021.+R+package+version+1.6.+0.+Ref.+Source. (Accessed: 3 December 2023).

Chen, S. *et al.* (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics*, 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTY560.

Divashuk, M.G. *et al.* (2014) 'Molecular Cytogenetic Characterization of the Dioecious Cannabis sativa with an XY Chromosome Sex Determination System', *PLOS ONE*, 9(1), p. e85118. Available at: https://doi.org/10.1371/journal.pone.0085118.

Dowling, C.A. *et al.* (2023) 'A FLOWERING LOCUS T ortholog is associated with photoperiod-insensitive flowering in hemp (Cannabis sativa L.)'. bioRxiv, p. 2023.04.21.537862. Available at: https://doi.org/10.1101/2023.04.21.537862.

ElSohly, M.A. and Slade, D. (2005) 'Chemical constituents of marijuana: The complex mixture of natural cannabinoids', *Life Sciences*, 78(5), pp. 539–548. Available at: https://doi.org/10.1016/j.lfs.2005.09.011.

Gao, C. *et al.* (2018) 'Genome-Wide Expression Profiles of Hemp (*Cannabis sativa* L.) in Response to Drought Stress', *International Journal of Genomics*, 2018, p. e3057272. Available at: https://doi.org/10.1155/2018/3057272.

Gao, S. *et al.* (2020) 'A high-quality reference genome of wild Cannabis sativa', *Horticulture Research*, 7(1), p. 73. Available at: https://doi.org/10.1038/s41438-020-0295-3.

Grassa, C.J. *et al.* (2018) 'A complete Cannabis chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content', *bioRxiv*, p. 458083. Available at: https://doi.org/10.1101/458083.

Guerriero, G. *et al.* (2017) 'Transcriptomic profiling of hemp bast fibres at different developmental stages.', *Scientific reports*, 7(1), p. 4961. Available at: https://doi.org/10.1038/s41598-017-05200-8.

Hussain, T. *et al.* (2021) 'Cannabis sativa research trends, challenges, and new-age perspectives', *iScience*, 24(12), p. 103391. Available at: https://doi.org/10.1016/j.isci.2021.103391.

Jain, A. and Tuteja, G. (2019) 'TissueEnrich: Tissue-specific gene enrichment analysis', *Bioinformatics*. Edited by J. Kelso, 35(11), pp. 1966–1967. Available at: https://doi.org/10.1093/bioinformatics/bty890.

Jiang, Y. *et al.* (2021) 'Physiological and transcriptome analyses for assessing the effects of exogenous uniconazole on drought tolerance in hemp (Cannabis sativa L.)', *Scientific Reports*, 11(1), p. 14476. Available at: https://doi.org/10.1038/s41598-021-93820-6.

Jones, P. *et al.* (2014) 'InterProScan 5: genome-scale protein function classification', *Bioinformatics*, 30(9), pp. 1236–1240. Available at: https://doi.org/10.1093/bioinformatics/btu031.

Joshi, C.J. *et al.* (2022) 'What are housekeeping genes?', *PLOS Computational Biology*. Edited by C. Kaleta, 18(7), p. e1010295. Available at: https://doi.org/10.1371/journal.pcbi.1010295.

Kanehisa, M., Sato, Y. and Morishima, K. (2016) 'BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences', *Journal of molecular biology*, 428(4), pp. 726–731. Available at: https://doi.org/10.1016/J.JMB.2015.11.006.

Kolde, R. (2019) 'pheatmap: Pretty Heatmaps'. Available at: https://cran.r-project.org/web/packages/pheatmap/index.html (Accessed: 4 September 2024).

Laverty, K.U. *et al.* (2019) ‘A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci.’, *Genome research*, 29(1), pp. 146–156. Available at: https://doi.org/10.1101/gr.242594.118.

Lawrence, M. *et al.* (2013) ‘Software for computing and annotating genomic ranges.’, *PLoS computational biology*. Edited by A. Prlic, 9(8), p. e1003118. Available at: https://doi.org/10.1371/journal.pcbi.1003118.

Livingston, S.J. *et al.* (2020) ‘Cannabis glandular trichomes alter morphology and metabolite content during flower maturation’, *The Plant Journal*, 101(1), pp. 37–56. Available at: https://doi.org/10.1111/TPJ.14516.

Lun, A.T.L., McCarthy, D.J. and Marioni, J.C. (2016) ‘A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor’, *F1000Research*, 5. Available at: https://doi.org/10.12688/F1000RESEARCH.9501.2.

Machado, F.B. *et al.* (2020) ‘Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (Glycine max) expression atlas’, *The Plant Journal*, 103(5), pp. 1894–1909. Available at: https://doi.org/10.1111/TPJ.14850.

Martin Morgan *et al.* (2023) ‘Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import’. Available at: https://bioconductor.org/packages/Rsamtools}.

McGarvey, P. *et al.* (2020) ‘De novo assembly and annotation of transcriptomes from two cultivars of Cannabis sativa with different cannabinoid profiles’, *Gene*, 762, p. 145026. Available at: https://doi.org/10.1016/j.gene.2020.145026.

McKernan, K.J. *et al.* (2020) ‘Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes’, *bioRxiv*, p. 2020.01.03.894428. Available at: https://doi.org/10.1101/2020.01.03.894428.

Medicinal Genomics (2019) ‘Jamaican Lion: The Industry’s First Comprehensive Cannabis Reference Genome’, *Medicinal Genomics*, 22 January. Available at: https://medicinalgenomics.com/jamaican-lion-cannabis-genome/ (Accessed: 20 August 2024).

de Meijer, E.P.M. *et al.* (2003) ‘The inheritance of chemical phenotype in Cannabis sativa L.’, *Genetics*, 163(1), pp. 335–46. Available at: https://doi.org/10.1093/genetics/163.1.335.

Mi, Y. *et al.* (2023) ‘Characterization and co-expression analysis of ATP-binding cassette transporters provide insight into genes related to cannabinoid transport in *Cannabis sativa* L.’, *International Journal of Biological Macromolecules*, 242, p. 124934. Available at: https://doi.org/10.1016/j.ijbiomac.2023.124934.

Patro, R. *et al.* (2017) ‘Salmon provides fast and bias-aware quantification of transcript expression’, *Nature Methods*, 14(4), pp. 417–419. Available at: https://doi.org/10.1038/nmeth.4197.

Pépin, N., Hebert, F.O. and Joly, D.L. (2021) ‘Genome-Wide Characterization of the MLO Gene Family in Cannabis sativa Reveals Two Genes as Strong Candidates for Powdery Mildew Susceptibility’, *Frontiers in Plant Science*, 12, p. 729261. Available at: https://doi.org/10.3389/fpls.2021.729261.

Plotly Technologies Inc (2015) ‘Collaborative data science.’ Montreal, QC. Available at: https://plot.ly.

Prentout, D. *et al.* (2020) ‘An efficient RNA-seq-based segregation analysis identifies the sex chromosomes of Cannabis sativa’, *Genome Research*, 30(2), pp. 164–172. Available at: https://doi.org/10.1101/gr.251207.119.

Ren, M. *et al.* (2019) ‘The origins of cannabis smoking: Chemical residue evidence from the first millennium BCE in the Pamirs’, *Science Advances*, 5(6), pp. 1391–1403. Available at: https://doi.org/10.1126/sciadv.aaw1391.

Richardson, N. *et al.* (2021) ‘Arrow: Integration to “Apache”’Arrow’’.

Skirycz, A. (2007) *Functional analysis of selected DOF transcription factors in the model plant Arabidopsis thaliana*. Universität Potsdam.

Stout, J.M. *et al.* (2012) ‘The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes.’, *The Plant journal : for cell and molecular biology*, 71(3), pp. 353–65. Available at: https://doi.org/10.1111/j.1365-313X.2012.04949.x.

Tang, Q. *et al.* (2023) ‘Transcriptomic and metabolomic analyses reveal the differential accumulation of phenylpropanoids and terpenoids in hemp autotetraploid and its diploid progenitor’, *BMC Plant Biology*, 23, p. 616. Available at: https://doi.org/10.1186/s12870-023-04630-z.

The UniProt Consortium (2023) ‘UniProt: the Universal Protein Knowledgebase in 2023’, *Nucleic Acids Research*, 51(D1), pp. D523–D531. Available at: https://doi.org/10.1093/nar/gkac1052.

Tian, F. *et al.* (2020) 'PlantRegMap: charting functional regulatory maps in plants', *Nucleic Acids Research*, 48(D1), pp. D1104–D1113. Available at: https://doi.org/10.1093/nar/gkz1020.

United Nations (2024) *Measuring Global Exports of Industrial Hemp Products: Insights from National Product Classifications*. UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT. Geneva: United Nations. Available at: https://doi.org/10.18356/9789213589281.

Van Der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9, pp. 2579–2605.

Wickham, H. (2016) 'ggplot2'. Available at: https://doi.org/10.1007/978-3-319-24277-4.

Yan, B. *et al.* (2023) 'Genome-Wide Identification, Classification, and Expression Analyses of the CsDGAT Gene Family in Cannabis sativa L. and Their Response to Cold Treatment', *International Journal of Molecular Sciences*, 24(4), p. 4078. Available at: https://doi.org/10.3390/ijms24044078.

Yeo, H.C. *et al.* (2022) 'Comparative Transcriptome Analysis Reveals Coordinated Transcriptional Regulation of Central and Secondary Metabolism in the Trichomes of Cannabis Cultivars', *International Journal of Molecular Sciences*, 23(15), p. 8310. Available at: https://doi.org/10.3390/ijms23158310.

Yin, M. *et al.* (2022) 'Proanthocyanidins Alleviate Cadmium Stress in Industrial Hemp (Cannabis sativa L.)', *Plants*, 11(18), p. 2364. Available at: https://doi.org/10.3390/plants11182364.

Zager, J.J. *et al.* (2019) 'Gene Networks Underlying Cannabinoid and Terpenoid Accumulation in Cannabis', *Plant Physiology*, 180(4), pp. 1877–1897. Available at: https://doi.org/10.1104/pp.18.01506.