

https://dspace.org/

Genome assembly, annotation and comparative analysis of Cannabis sativa Swazi Gold (SG)

Glen, Tereka

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/50282
Glen, Tereka. (2025). Genome assembly, annotation and comparative analysis of Cannabis sativa Swazi Gold (SG) [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/50282

Downloaded from a DSpace repository. https://dspace.org/



Genome assembly, annotation and comparative analysis of *Cannabis sativa* Swazi Gold (SG)

by

Tereka Glen

(2094806)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Angela Botes

Co-supervisor: Prof Pieter De Maayer

February 2026

Declaration

I, Tereka Glen, (student number: 2094806) am a student registered for the degree of Master of Science (by dissertation) at the University of Witwatersrand in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the degree is my unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.



Signature

18 February 2026

Date

Wits University Faculty of Science post-graduate student AI declaration

I understand that the use of generative AI tools (such as ChatGPT or similar) without explicitly declaring such use constitutes a form of plagiarism and is classified by Wits University as academic misconduct.

I declare that in the course of conducting the research towards my degree or in the preparation of this thesis/dissertation/research report (select one by marking with an X):

I **did not** make use of generative AI tool ☐

I **did** make use of generative AI tools for the following (tick all that apply):

1. Idea Generation (research problem/design, hypothesis) ☐
2. Sourcing Related Work (summarising, identifying sources) ☐
3. Methods and Experiment Design (experiment setup, model tuning) ☒
4. Data Analysis (presentation, coding, interpretation) ☒
5. Theoretical Development (theorem proving, conceptual analysis) ☐
6. Code Development (generating algorithms, writing scripts) ☐
7. Presentation (rendering graphics, formatting) ☐
8. Editing (grammar, readability) ☐
9. Writing (text generation, document structuring) ☐
10. Citation Formatting (structuring, organising) ☐

If other uses were involved, please specify below:

Generative AI tool used (list all)	**Used for?**

If generative AI tools were used as an integral part of the experimental design or in the direct execution of my research, I confirm that details of this use are clearly outlined in the relevant experimental/methodology chapters of my thesis/dissertation/research report.

Student number: 2094806

Candidate signature: 

Date: 31.10.2025

Preface

Cannabis sativa L. (Linnaeus) is an annual herbaceous plant that produces 570 phytochemicals, with 125 of these being characterised as cannabinoids. Cannabinoids have demonstrated medicinal potential in the treatment of multiple conditions including glaucoma, sleeplessness, depression, Alzheimer's disease, neuralgia and multiple sclerosis. *Cannabis sativa* plants that have adapted to the environment of a specific geographic region are referred to as a "landrace"; these landraces differ in their cannabinoid profiles and production levels compared to commercial cultivars. Here we have sequenced and assembled the genome of *C. sativa* landrace Swazi Gold (SG), which originates in the Kingdom of Eswatini and represents the first genome of an African landrace. Subsequently we undertook genome annotation and comparative genomic analyses to gain a better understanding of the genetics underlying cannabinoid biosynthesis in an African landrace, which could in turn be exploited in various therapeutic and medicinal applications.

In Chapter, 1 the relevant literature available regarding the taxonomy, history and morphology of *C. sativa* is provided in a comprehensive review. This is followed by an overview on its medicinal applications, cannabinoid biosynthesis and the therapeutic applications of the major cannabinoids. Additionally, current knowledge on plant genome assembly, focusing on *C. sativa* genome research and its gaps is discussed.

In Chapter 2, the genome of *C. sativa* SG was sequenced using short-read and long-read approaches. Multiple *de novo* assembly strategies were evaluated to produce a high-quality draft genome assembly coupled with reference-based refinement and polishing. Subsequently, the refined draft genome underwent structural and functional annotation.

In Chapter 3, a comparative genomic analysis of the draft genome of *C. sativa* SG and 13 publicly available *C. sativa* genomes of *C. sativa* (and one additional *C. sativa* genome assembled in house) was performed, with specific focus on their cannabinoid biosynthetic and secondary metabolite loci.

Finally, a summary highlights key findings of this research and provides recommendations on future work that should be undertaken to further understand the evolutionary lineage and biosynthetic pathways of *C. sativa* to ultimately explore the therapeutic applications of this plant.

Dedication

I dedicate this work to my mother, father, and brother for their unconditional love, support, and their countless sacrifices that have carried me to this milestone.

I also dedicate this work to my beloved partner, whose support made this achievement possible.

Acknowledgements

I would like to express my deepest gratitude to God for blessing me with this opportunity.

I am eternally grateful to my supervisor, Dr. Angela Botes, for her invaluable guidance, encouragement and steadfast support throughout this project. I have learned, and continue to learn, so much from you. The skills and knowledge I have gained under your mentorship will remain with me throughout my career and life.

I would also like to sincerely thank my co-supervisor, Prof. Pieter de Maayer, for his insightful feedback, technical expertise and mentorship during this journey. Thank you for providing computational resources and, for taking the time to teach me and expand my bioinformatics skills.

Thank you to The ROOT and the University of the Witwatersrand for the Postgraduate Merit Award, for generously funding this research and enabling me to complete this project.

Thank you to my colleagues in the Botes Lab for their constant support and for creating such a collaborative environment. With a special mention for Mr Dylan Moodley for his unwavering guidance, support and for always believing in me.

I am so immensely thankful to my brother, mother and father for their unwavering belief in me, their unshakeable support and their encouragement. My mother, for reminding me to find strength in difficult times and being a source of inspiration. My father, instilling in me the belief that I can achieve anything I my mind to. My brother, for bringing light into my life, having you closer has been a constant source of comfort.

Finally, I would like to thank my best friend and partner Mr Mazwi Meyiwa, for his steadfast support, encouragement and patience. You have been my rock, keeping me grounded when the world felt unsteady, and my source of strength throughout this journey.

Table of Contents

Declaration 2

Preface 4

Dedication 5

Acknowledgements 6

Table of Contents 7

List of Figures 10

List of Tables 12

List of Symbols and Abbreviations 14

Chapter one: Literature Review 17

1.1. Introduction 18

1.2. Environmental origins and global distribution 18

1.2. Plant morphology and life cycle 20

1.3. History of the use of *Cannabis* 21

1.4. Cannabinoids 22

1.5. Cannabinoid biosynthesis in *C. sativa* 23

1.6. Medical applications of *C. sativa* cannabinoids 25

1.6.1. Medical relevance of CBD 26

1.6.2. Medical relevance of Δ^9 -THC 26

1.6.3. Medical relevance of CBC 27

1.6.4. Medical relevance of CBG 27

1.7. Applications of olivetol and olivetolic acid 28

1.8. Plant genome assembly 29

1.9. Genomic studies of *C. sativa* 30

1.9.1. Genome repetitiveness and ploidy in *Cannabis* 31

1.10. Single nucleotide variants in *C. sativa* 33

1.11. Purpose of this study 34

1.12. Aim and objectives 34

1.13. References 35

Chapter Two: Sequencing, assembly and annotation of the *Cannabis sativa* Swazi Gold genome 52

2.1. Abstract 53

2.2. Introduction 53
2.3. Methods and Materials 55
2.3.1. Plant material 55
2.3.2. DNA Extraction 55
2.3.3. Library preparation and sequencing 57
2.3.4. Genomic sequence quality control and trimming 58
2.3.4. Short-read *de novo* genome assembly 58
2.3.5. Evaluation of *de novo* genome assemblies 59
2.3.6. Hybrid assembly (Illumina short-read and PacBio long-read) 60
2.3.7. Reference-guided assembly 62
2.3.8. Refinement of the *C. sativa* SG RaGOO-guided genome assembly 63
2.3.9. Genome annotation 64
2.4. Results and Discussion 66
2.4.4. DNA extraction 66
2.4.5. Genomic sequence quality control 66
2.4.6. *de novo* Genome assembly 68
2.4.7. Reference-guided assembly 71
2.4.8. Refinement of the draft *C. sativa* SG genome assembly 72
2.4.9. Comparison of the final SG genome assemblies 73
2.4.10. Selection and evaluation of SG final draft genome 79
2.5. Conclusion 80
2.6. References 81
Chapter Three: Comparative genomic analysis of *Cannabis sativa* Swazi Gold 91
3.1. Abstract 92
3.2. Introduction 92
3.3. Methods and Materials 93
3.3.1. Structural and functional annotation of available genomes 93
3.3.2. Single-copy orthologue identification and core genome phylogeny construction 95
3.3.3. BLAST analysis of cannabinoid biosynthetic genes in *C. sativa* 95
3.3.4. Biosynthetic gene cluster (BGC) identification 97
3.4. Results and Discussion 97
3.4.1. Comparison of the annotated *C. sativa* genomes 97

3.4.2. Single-copy orthologue identification and core genome phylogeny construction 100
3.4.3. Analysis of cannabinoid biosynthetic genes in *C. sativa* 103
3.4.4. Biosynthetic gene cluster analysis 114
3.5. Conclusion 116
3.6. References 117
Chapter Four: Summary 129
Chapter 5: Appendix 132

List of Figures

Figure 1.1: Geographical distribution of legal *Cannabis sativa* cultivation**Error! Bookmark not defined.**

Figure 1.2: Morphological structure of male and female *Cannabis* plants (Fidan *et al.,* 2023). ..... **Error! Bookmark not defined.**

Figure 1.3: Alkyl type cannabinoid with monoterpene moiety and pentyl side chain (Kanabus *et al.,* 2021). ..... **Error! Bookmark not defined.**

Figure 1.4: Biosynthesis pathway of cannabinoids. ..... **Error! Bookmark not defined.**

Figure 1.5: Formation of olivetolic acid (OLA) and olivetol. **Error! Bookmark not defined.**

Figure 2.1: Proportion of tRNAs by amino acid within the Rebaler reference-guided (Pie chart 1) and the RaGOO reference-guided (with the additional RaGOO contigs) *Cannabis sativa* Swazi Gold assemblies. .....

Error! Bookmark not defined.

Figure 2.2: Ribosomal RNA (rRNA) composition of the Rebaler reference-guided (Pie chart 1) and the RaGOO reference-guided (with additional RaGOO contigs) (Pie chart 2) *Cannabis sativa* Swazi Gold assemblies. ..... **Error! Bookmark not defined.**

Figure 3.1: Core Maximum likelihood (ML) phylogeny of *Cannabis sativa* cultivars and landraces. 101

Figure 3.2: Neighbour-Joining (NJ) phylogenetic tree of acetyl-CoA carboxylase (ACCase) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. .....**Error! Bookmark not defined.**

Figure 3. 3: Neighbour-Joining (NJ) phylogenetic tree of hexanoyl-CoA synthetase (HCAS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes.. 105

Figure 3.4: Neighbour-Joining (NJ) phylogenetic tree of tetraketide synthase (OLS / TKS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. 107

Figure 3.5: Neighbour-Joining (NJ) phylogenetic tree of olivetolic acid cyclase (OAC), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. 108

Figure 3. 6: Neighbour-Joining (NJ) phylogenetic tree of 1-deoxy-D-xylulose-5-phosphate synthase (DXS) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes.. 109

Figure 3. 7: Neighbour-Joining (NJ) phylogenetic tree of geranyl pyrophosphate synthase (GPPS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. .110

Figure 3. 8: Neighbour-Joining (NJ) phylogenetic tree of geranylpyrophosphate olivetolate geranyltransferase (GOT) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. 111
Figure 3. 9: Neighbour-Joining (NJ) phylogenetic tree of cannabidiolic acid synthase (CBDAS), cannabichromenic acid synthase (CBCAS) and tetrahydrocannabinolic acid synthase (THCAS) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. 113
Appendix Figure A1: Quality of the *Cannabis sativa* Swazi Gold raw Illumina short-read data 132
Appendix Figure A2: Quality of the *Cannabis sativa* Swazi Gold Illumina short-read data post trimming using TrimGalore v. 0.6.7 with a Phred score cutoff of Q20. 133

List of Tables

Table 2.1: MultiQC v 1.12 and FAstQC v 0.11.9 Illumina and PacBio raw sequence data analysis results ..... **Error! Bookmark not defined.**

Table 2.2: Illumina short-read sequence data results after trimming at Phred score cutoffs of Q20 and Q25 using TrimGalore v. 0.6.7 with Cutadapt v. 3.5 **Error! Bookmark not defined.**

Table 2.3: Genome metrics of assemblies produced on the Q20 trimmed Illumina short-read data with AbySS v. 4.4., MEGAHIT v 1.2.9 and SGA v. 0.10.15.**Error! Bookmark not defined.**

Table 2.4: Optimal hybrid *de novo* assemblies produced using AbySS v. 4.4., MEGAHIT v. 1.2.9 and SGA v. 0.10. ..... **Error! Bookmark not defined.**

Table 2.5: Assemblies generated using two reference-guide assembly strategies RaGOO v. 1.1 and Rebaler from the MEGAHIT hybrid genome assembly, at default settings. .....**Error! Bookmark not defined.**

Table 2.6: Comparison of the Rebaler assembly, refined RaGOO and refined additional RaGOO contig assemblies. ..... **Error! Bookmark not defined.**

Table 2.7: Comparison of MOSGA v. 2.1.5 annotation results for the Rebaler assembly, refined RaGOO and refined additional RaGOO contig assemblies for *Cannabis sativa* Swazi Gold..... **Error! Bookmark not defined.**

Table 3.1: Overview of fifteen comparator *Cannabis sativa* genomes with chemotype and geographic location data 102

Table 3. 2: Overview of *Cannabis sativa* cannabinoid biosynthetic proteins retrieved from the NCBI database 104

Appendix A Table 1: *Cannabis sativa* genomes available on the NCBI database for different landraces/cultivars. 145

Appendix A Table 2: Annotation results of fifteen *Cannabis sativa* genomes used in this study. 147

Appendix A Table 3: Comparison of the composition of repetitive elements in the Rebaler reference-guided *Cannabis sativa* Swazi Gold genome to the composition repeat element range for the fourteen annotated *Cannabis sativa* genomes, determined using RepeatMasker. 150

Appendix A Table 4: The rRNA predication results of *Cannabis sativa* genomes available on the NCBI database for different landraces/cultivars, using Barrnap v. 0.9. 151

Appendix A Table 5: Local BLAST results for the identification of cannabinoid biosynthetic genes in the annotated *Cannabis sativa* genomes.....152
Appendix A Table 6: Biosynthetic gene cluster (BGC) identification results for all fiffteen annotated *Cannabis sativa* genomes using PlantiSMASH.....154

List of Symbols and Abbreviations

ACCase: Acetyl-CoA Carboxylase

AD: Alzheimer's Disease

AIDS: Acquired Immunodeficiency Syndrome

ALS: Amyotrophic Lateral Sclerosis

AP: Aromatic Prenyltransferase

BGCs: Biosynthetic Gene Clusters

bp: Base Pair

BWA: Burrows–Wheeler Aligner

CBC: Cannabichromene

CBCA: Cannabichromenic Acid

CBCAS: Cannabichromenic Acid Synthase

CBD: Cannabidiol

CBDA: Cannabidiolic Acid

CBDAS: Cannabidiolic Acid Synthase

CBDVA: Cannabidivarinic Acid

CBG: Cannabigerol

CBGA: Cannabigerolic Acid

CBGAS: Cannabigerolic Acid Synthase

CB: Cannabinoid

CNS: Central Nervous System

CNV: Copy number variation

CPU: Central Processing Unit

COX: Cyclooxygenase

DBG: De Bruijn Graph

DBG2OLC: De Bruijn Graph to Overlap-Layout-Consensus

DNA: Deoxyribonucleic Acid

DXS: 1-Deoxy-D-xylulose-5-phosphate Synthase

EBP: Earth BioGenome Project

EMBL-EBI: European Molecular Biology Laboratory's European Bioinformatics Institute

gDNA: Genomic DNA

GOT: Geranylpyrophosphate Olivetolate Geranyltransferase

GPPS: Geranyl Pyrophosphate Synthase

GIT: Gastrointestinal Tract

GPP: Geranyl Pyrophosphate

HCASs: Hexanoyl-CoA Synthetase

HIV: Human Immunodeficiency Virus

LOX: Lipoxygenase

LTR: Long Terminal Repeat Transposons

Mbp: Megabase Pair

MEP: Methylerythritol-4-Phosphate

MIC: Minimum Inhibitory Concentration

ML: Maximum-likelihood

MOSGA: Modular Open-Source Genome Annotator

m: Metre

NCBI: National Centre for Biotechnology Information

NGS: Next Generation Sequencing

NJ: Neighbour-joining

OAC: Olivetolic Acid Cyclase

OLA: Olivetolic Acid

OLS: Olivetol Synthase

ONT: Oxford Nanopore Technologies

PacBio: Pacific Biosciences

PD: Parkinson’s Disease

PE: Paired-End

PKS: Polyketide Synthase

PNS: Peripheral Nervous System

PTSD: Post-Traumatic Stress Disorder

rRNAs: Ribosomal RNA Elements

SCOs: Single-Copy Orthologous Genes

SdBG: Succinct De Bruijn Graph

SG: Swazi Gold

SMRT: Single Molecule Real-Time

SNV: Single Nucleotide Variant

SNP: Single Nucleotide Polymorphism

TGS: Third Generation Sequencing

THCAS: Tetrahydrocannabinolic Acid Synthase

TKS: Tetraketide Synthase

tRNAs: Transfer RNA Elements

UHDRS: Unified Huntington's Disease Rating Scale

USA: United States of America

ZMWs: Zero-Mode Waveguides

Δ^9-THC: Δ9-Tetrahydrocannabinol

Δ^9-THCA: Δ9-Tetrahydrocannabinolic Acid

Δ^9-THCVA: Δ9-Tetrahydrocannabivarinic Acid

Chapter one: Literature Review

1.1. Introduction

Cannabis sativa L. is an annual herbaceous plant belonging to the family *Cannabaceae* and order Rosales (Radwan *et al.,* 2021). It is widely accepted that the genus *Cannabis* comprises of a single (monotypic) species, *C. sativa*, with polymorphic subspecies namely *indica, sativa* and *ruderalis* (McPartland, 2018; Sawler *et al.,* 2015; Small & Cronquist, 1976). However, some scientists argue that the genus is polytypic (Hillig, 2005; Steel *et al.,* 2023), incorporating three distinct species, namely *C. sativa, C. indica Lam* (Lamarck) and *C. ruderalis*, with each displaying significant phenotypic differences (Sawler *et al.,* 2015).

Cannabis sativa thrives in a wide variety of environments, ranging from sea level and temperate zones to alpine foothills (Radwan *et al.,* 2021). Its growth outside of the agricultural setting predominantly includes habitats created by or associated with humans, such as disturbed land, landfills, farmland, roadsides, vacant lots, railways, trails and exposed riverbanks (Clarke & Watson, 2007). This plant has the potential to grow up to five metres in height under optimal conditions (Braich *et al.,* 2020).

Cannabis sativa is dioecious, having both male and female plants (Braich *et al.,* 2020; Hurgobin *et al.,* 2020). It is a diploid, comprising nine autosomes and one pair of sex chromosomes (X and Y) (Braich *et al.,* 2020). The estimated diploid genome size of the female and male plants has been determined using flow cytometry to be 1,636 ± 7.2 and 1,683 ± 13.9 Mbp, respectively (Braich, 2020; Hurgobin *et al.,* 2020).

1.2. Environmental origins and global distribution

Cannabis originated in Central Asia, approximately 12,000 years ago, making it one of the oldest cultivated crops (Bradshaw *et al.,* 1981; Long *et al.,* 2017). As knowledge of the advantages associated with this plant increased, *Cannabis* cultivation spread from Asia across the globe (Fidan *et al.,* 2023). By the early nineteenth century, it was introduced into western medicine and is currently cultivated legally and illegally on every continent except for Antarctica (Figure 1.1) (Lynch *et al.,* 2016; van Bakel *et al.,* 2011; Zuardi, 2006). Due to the extensive geographical and ecological range of *C. sativa*, outdoor populations can be found in a variety of habitats including the Arctic circle, with tropical populations being very rare (Clarke & Merlin, 2013). When a *Cannabis* plant is indigenous to a specific geographic region and has adapted to the environmental conditions in that region it is referred to as a "landrace"

(Sawler *et al.,* 2015). Each landrace has unique characteristics including cannabinoid profile, growth patterns and morphology (Braich *et al.,* 2020). The diverse phytochemical composition, morphological and evolutionary characteristics of *C. sativa* have led to multiple debates on its taxonomic status and origin of domestication. Currently, there is no consensus regarding the exact location of its origin (Hillig & Mahlberg, 2005; Hillig, 2005).

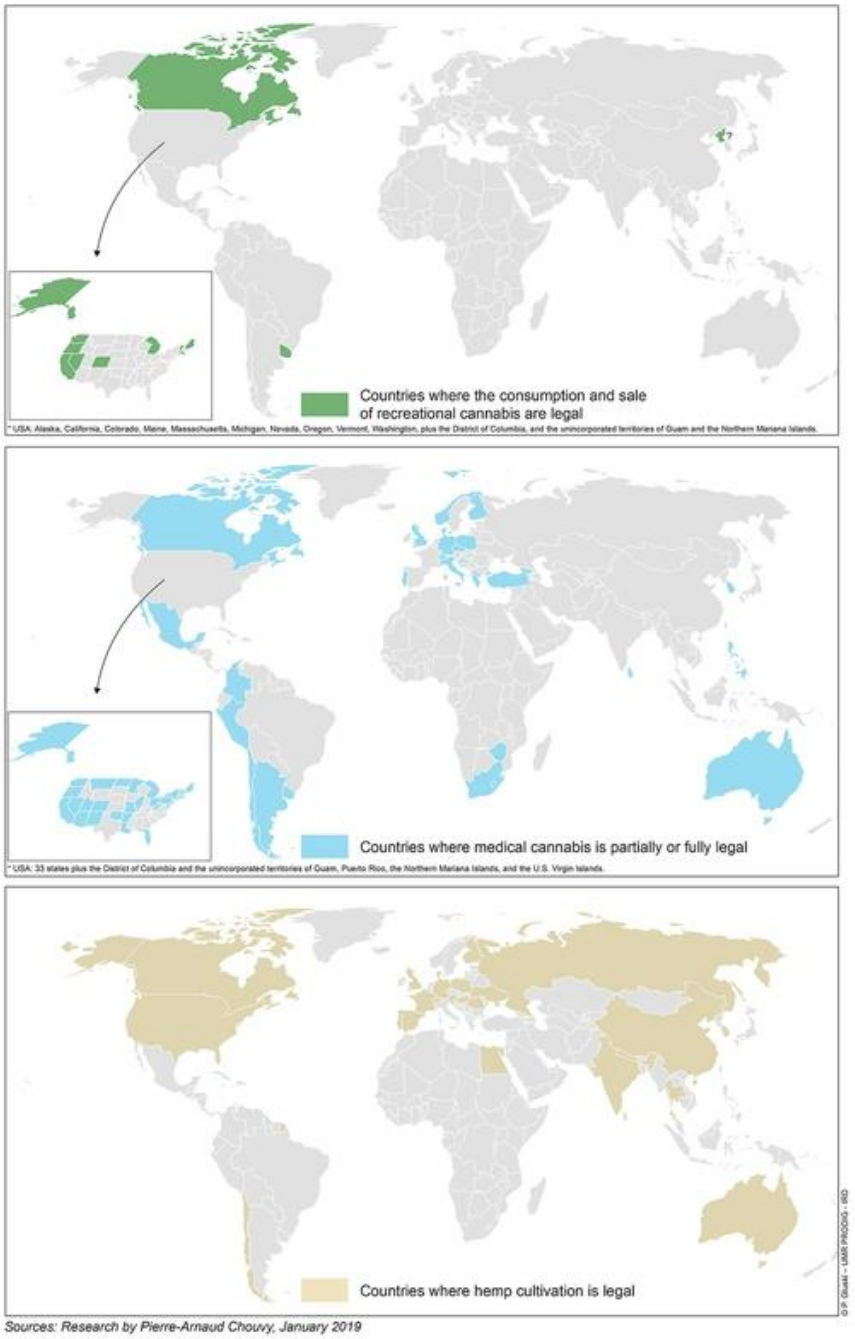


Figure 1.1: Geographical distribution of legal *Cannabis sativa* cultivation based on its application as of January 2019 (Image from Chouvy, 2019).

1.2. Plant morphology and life cycle

Cannabis is propagated from seed and considered an annual herb, completing one life cycle within one year (Radwan *et al.,* 2021; Ranalli, 1999). The life cycle is comprised of four defined stages: germination, a seedling stage, vegetative stage and a flowering stage (Cervantes, 2007). These stages differ in environmental, nutritional and sunlight requirements (Radwan *et al.,* 2021).

The first stage, germination, begins when ripe seeds (that have fallen to the ground or have been planted) are provided with ideal growth conditions (Ranalli, 1999). These conditions include sufficient water, nutrients and sunlight. During germination the seedling, consisting of cotyledons (seed leaves), sprouts from the seed coat (Cervantes, 2007). The vegetative stage consists of substantial increase in plant biomass when the daylight hours per day increase (Ranalli, 1999). This phase also includes plant hormone changes, the formation of leaves and trichomes, root growth for increased nutrient and water uptake and the upregulation of genes important to this phase-transition (Spitzer-Rimon *et al.,* 2022). To complete its life cycle *Cannabis* must flower and produce seeds (Cervantes, 2007). After four weeks of vegetative growth male and female pre-flowers are produced, leading to the flowering stage (Spitzer-Rimon *et al.,* 2022). Flowering occurs when the plant is exposed to day lengths of less than 12-14 hours and long nights of 10-12 hours (Clarke & Watson, 2007). The flowering stage is characterised by a change in plant biomass to floral induction, including male and female flowers, and increased resin production. Pollen grains from male flowers are transported via air currents, which facilitates contact with female flowers. Pollination and pollen shedding from the male plants usually lasts between two to four weeks (Cervantes, 2007; Ranalli, 1999). Following fertilisation, the single seeds in the female flowers ripen in three to eight weeks. Ripe seeds are either harvested or fall to the ground where they will germinate and the plants life cycle repeats (Cervantes, 2007).

Cannabis sativa plant produces palmate leaves made up of five to seven leaflets which taper at both ends and have serrated margins (Farag & Kayser, 2017; Happyana *et al.,* 2013; Huchelmann *et al.,* 2017). It is rich in glandular trichomes which cover the stems, bracts and leaves of the plant. In female plants these glandular trichomes are mostly found at the aerial centre of the plant and enclose and synthesize secondary metabolites including terpenoids and phytocannabinoids (Andre *et al.,* 2016). This cannabinoid production is impacted by sunlight, changing with increased sunlight exposure (Hesami *et al.,* 2023). The distinction between male and female plants can be made during plant growth and flowering (Fidan *et al*, 2023). During

plant growth the male plant has thick, large stalks and the buds are replaced by small flat round nodes located where the stem meets the branches (Fidan *et al*, 2023; Sawler *et al.,* 2015). Female plant growth produces buds with a higher leaf count, located predominantly on the branches and form fruit during the flowering stage (Figure 1.2). The flowers produced by a male plant do not present with petals or panicles and have five anthers and five tepals that are yellowish in colour (Bonini *et al.,* 2018). Female flowers sprout in the axils and terminally, from which a single fruit is produced per flower.

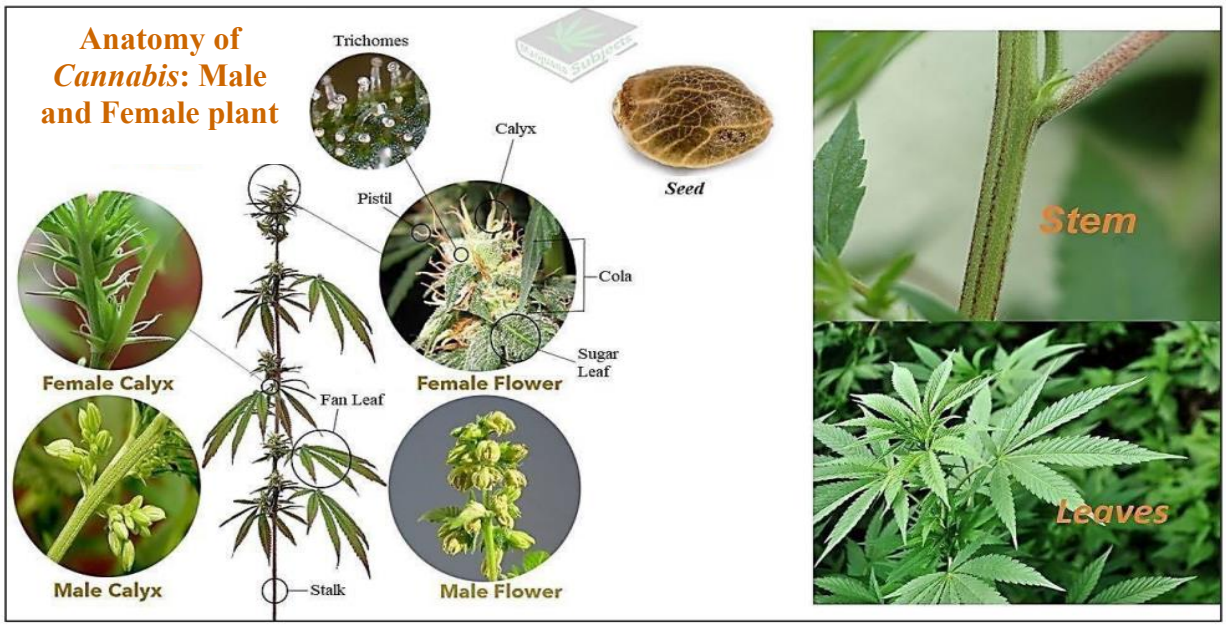


Figure 1.2: Morphological structure of male and female *Cannabis* plants (Image from Fidan *et al.,* 2023).

1.3. History of the use of *Cannabis*

Historically the *Cannabis* plant was used for food, fibre, hemp seeds and medicinal applications (Zuardi, 2006). The first known record of *Cannabis* used for these purposes dates back 5,000 years in China (Bonini *et al.,* 2018). *Cannabis* has also been recorded in ancient books written about plant medicine as a therapeutic herb, where it was mostly used as a component of a tea (Hazenkamp *et al.,* 2010).

Following its introduction into Western medicine, *Cannabis* was used to treat a variety of conditions, including asthma, epilepsy, fatigue, glaucoma, sleeplessness, labour pains, menstrual cramps, migraine pain and nausea (Batchelder, 2004; Zuardi, 2006). However, as a

results of the lack of quality control of *Cannabis* and its products, preparing a standard medical dose was difficult, and often resulted in patients receiving doses that were either too low and ineffective, or too high, leading to serious side effects (Russo, 2007). These drawbacks reduced the use of medicinal *Cannabis*, leading to its removal in most Western countries in the beginning of the twentieth century (Hazenkamp *et al.,* 2010; Zuardi, 2006).

1.4. Cannabinoids

Cannabis research has experienced an uptick following the identification of the endogenous cannabinoid system, a neuromodulatory system responsible for partially regulating cognitive and emotional processes in the human central nervous system (CNS) (Grotenhermen, 2005). Cannabinoids are compounds that interact and associate with mammalian cannabinoid receptors and evoke biological effects through this interaction (Duggan, 2021). There are three groups of cannabinoids: phytocannabinoids (found in plants), endocannabinoids (endogenously produced in humans and animals) and synthetic cannabinomimetics (synthetic compounds that produce antagonistic effects towards cannabinoid receptors) (Shevyrin & Morzherin, 2015).

Cannabis sativa produces 570 phytochemicals with 125 of these being characterised as phytocannabinoids, and the remaining chemicals being characterised as either flavonoids, alkaloids or terpenes (Pellati *et al.,* 2018). Cannabinoids produced by *C. sativa* are predominantly alkyl type cannabinoids, carrying a monoterpene moiety (C_{10}) with a pentyl side chain (C_5) (Figure 1.3) (Hanuš *et al.,* 2016). The most abundant cannabinoids present in *C. sativa* includes the psychotropic Δ^9-tetrahydrocannabinol (Δ^9-THC), and the non-psychotropics cannabidiol (CBD), cannabichromene (CBC), and cannabigerol (CBG) (ElSohly *et al.,* 2017; Happyana & Kayser, 2016). Psychotropic refers to a substance that affects mental and psychological functions leading to changes in mood, feelings, behaviour, awareness and thoughts (ElSohly *et al.,* 2017).

Based on their cannabinoid profile, *C. sativa* plants are divided into five chemotypes (Appendino *et al.,* 2011; Hartsel *et al.,* 2016; Tomas & ElSohly, 2015). Chemotype I consist of drug-type plants predominantly producing Δ^9-THC-type cannabinoids. Chemotypes III and IV comprises fibre-type plants with high non-psychoactive cannabinoid levels and low psychoactive cannabinoid levels (Appendino *et al.,* 2011). By contrast, Chemotype II plants have intermediate characteristics of both fibre-type and drug-type plants. And lastly,

chemotype V includes fibre-type plants which produce almost no cannabinoids (Appendino *et al.,* 2011; Tomas & ElSohly, 2015). Drug-type plants are used for recreational and medicinal applications, while fibre-type plants are used for food or textile applications (Taura *et al.,* 2007). The Δ^9-THC levels of *Cannabis* plants classified as industrial hemp with fibre-type chemotypes are limited, typically ranging between 0.2-1% (Salentijn *et al.,* 2015).

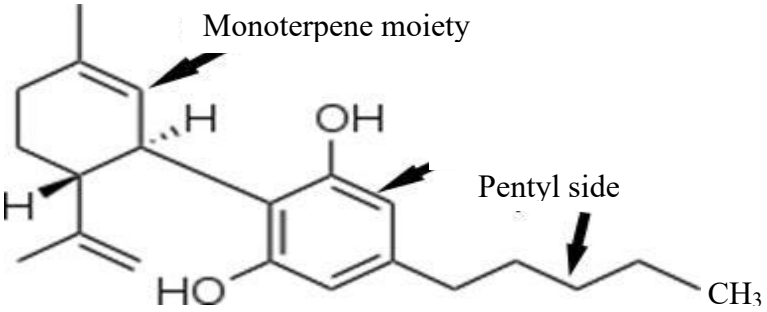


Figure 1.3: Alkyl type cannabinoid with monoterpene moiety and pentyl side chain (Image from Kanabus *et al.,* 2021).

1.5. Cannabinoid biosynthesis in *C. sativa*

Cannabinoids are synthesized in the glandular trichomes of *C. sativa* from fatty acids and isoprenoid precursors (Figure 1.4) (Andre *et al.,* 2016; Gagne *et al.,* 2012). Biosynthesis begins with a polyketide synthetic pathway in the cytosol, which produces olivetolic acid (OLA) and divarinic acid (Dewick, 2009; Stout *et al.,* 2012). In this pathway, tetraketide synthase (TKS) catalyses the condensation of one hexanoyl-CoA (to produce OLA) or one butanoyl-CoA (to produce divarinic acid) with three malonyl-CoA molecules, producing a linear tetraketide intermediate (3,5,7-trioxododecanoyl-CoA) (Gagne *et al.,* 2012). This intermediate will undergo a C2 → C7 aldol condensation performed by olivetolic acid cyclase (OAC) to form Olivetolic acid and divarinic acid as the major products (Taura *et al.,* 2007). These products then serve as the precursors from which all major and minor cannabinoids are derived through the methylerythritol-4-phosphate (MEP) pathway (Kim *et al.,* 2022).

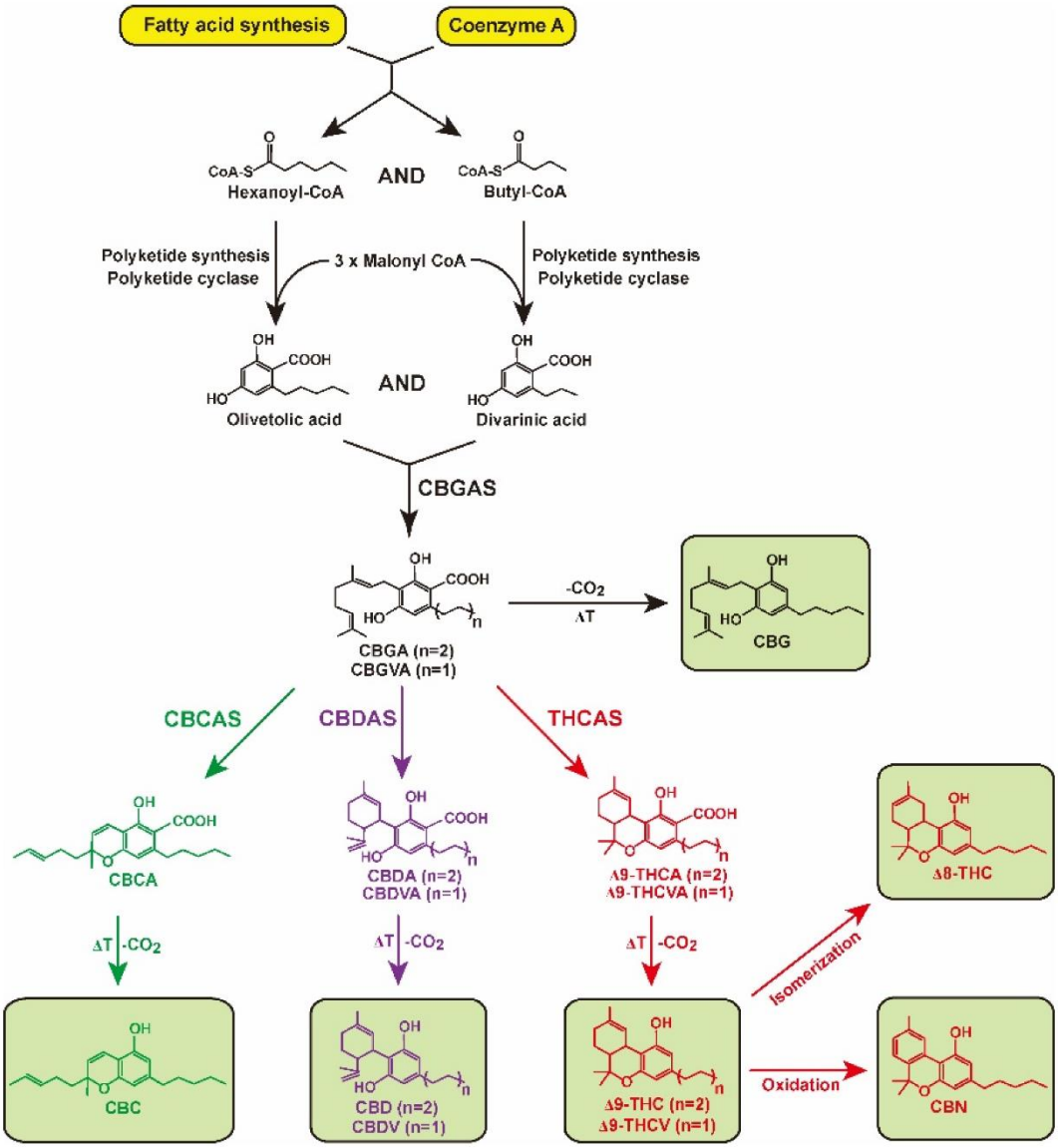


Figure 1.4: Cannabinoid biosynthetic pathway in *Cannabis sativa* resulting in the production of major cannabinoids. Hexanoyl-CoA and butyl-CoA derived from fatty acid metabolism are used for the biosynthesis of olivetolic and divarinic acid (Image from Kim *et al.,* 2022). Through the action of cannabigerolic acid synthase (CBGAS) olivetolic and divarinic acids are converted to cannabinoid acids Cannabigerolic acid (CBGA), Cannabigerovarinic acid (CBGVA), cannabichromenic acid (CBCA), cannabidiolic acid (CBDA), cannabidivarinic acid (CBDVA), Δ^9-tetrahydrocannabinolic acid (Δ^9-THCA) and Δ^9-tetrahydrocannabivarinic acid (Δ^9 -THCVA). After heat exposure these acids are decarboxylated and converted to Cannabigerol (CBG), cannabichromene (CBC), cannabidiol (CBD), CBDV, Δ^9-tetrahydrocannabinol (Δ^9-THC) and Δ^9-tetrahydrocannabivarinol Δ^9-THCV). This study focused on the identification of these major cannabinoid biosynthetic genes in *C. sativa* Swazi Gold.

The MEP pathway generates isopentenyl diphosphate and dimethyl allyl diphosphate (Xie *et al.,* 2023). These compounds are combined to form geranyl pyrophosphate (GPP), providing the monoterpene fraction needed for the synthesis of cannabinoids (Gulck & Moller, 2020). In the presence of GPP, OLA and divarinic acid are prenylated (the addition of a hydrophobic molecule to a biomolecule or protein) to form cannabigerolic acid (CBGA) through the action of an aromatic prenyltransferase (Cannabigerolic acid synthase - CBGAS) (Luo *et al.,* 2019). CBGA undergoes oxidative cyclization for the synthesis of cannabichromenic acid (CBCA), cannabidiolic acid (CBDA), cannabidivarinic acid (CBDVA), Δ^9-tetrahydrocannabinolic acid (Δ^9-THCA), and Δ^9-tetrahydrocannabivarinic acid (Δ^9-THCVA) (Marks *et al.,* 2009; Taura *et al.,* 1996). The resultant cannabinoids are stored in the resin cavities of the plant in their carboxylic acid forms. These forms are unstable and readily decarboxylated when exposed to heat or light, resulting in their conversion to CBG, CBC, CBD, cannabidivarin (CBDV), Δ^9-THC and Δ^9-tetrahydrocannabivarin (Δ^9-THCV), respectively (Tan *et al.,* 2018).

1.6. Medical applications of *C. sativa* cannabinoids

The endocannabinoid system partially regulates emotional and cognitive processes in the CNS and has been linked to neurological, mood and behavioural disorders (Capodice & Kaplan, 2021). This system is also present throughout the peripheral nervous system (PNS) and has multiple modulators, which include synthetic antagonists, agonists of the cannabinoid (CB) receptors and inhibitors of endocannabinoid degradation (Grotenhermen, 2005). Cannabinoid receptors are distributed in the CNS and various peripheral tissues including the immune system, gastrointestinal tract (GIT), arteries, reproductive tract, heart and lungs (Grotenhermen, 2005). Cannabinoid receptor agonists have been identified to have therapeutic properties including muscle relaxation, mood improvement, analgesia, anti-inflammation, immunosuppression, appetite stimulation, neuroprotection and antineoplastic effects (Grotenhermen, 2003; Grotenhermen, 2005; Lowe *et al.,* 2021). Current clinical research on these CB receptors is focused on their application for the treatment of neurological disorders, chronic pain, obesity and nicotine addiction (Akbas *et al.,* 2008; Consroe, 1998; Manzanares *et al.,* 2006; Tucci *et al.,* 2006).

The interaction of *C. sativa* cannabinoids with this endocannabinoid system highlights their potential for the treatment of a vast number of illnesses, these include Alzheimer's disease (AD), depression, glaucoma, multiple sclerosis, neuralgia, sleeplessness, as well as the

alleviation of symptoms of diseases such as cancer, human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS) (Abrams *et al.,* 2007; ElSohly *et al.,* 2017; Liang *et al.,* 2004).

1.6.1. Medical relevance of CBD

Cannabidiol (CBD), a non-psychrotropic cannabinoid, has potential in treating a range of inflammatory conditions including cancer tumour development and progression, arthritis, colitis and inflammation in the gastrointestinal tract (Freeman *et al.,* 2019; Grivennikov, 2010; Khalsa *et al.,* 2022). Most cancers originate through inflammation, infection or chronic irritation, with the microenvironment of the tumour being largely regulated by inflammatory cells. This microenvironment plays important roles in the neoplastic process, proliferation, survival and migration of tumour cells. CBD has demonstrated proapoptotic and antiproliferative properties, modulating tumorigenesis in different cancer types such as breast, colon and lung cancers (Namdar & Koltai, 2018).

Infection caused by *Clostridioides difficile* is one of the leading-causes of hospital-acquired colitis and diarrhoea (Lofgren *et al.,* 2014). *Clostridioides difficile* toxin A damages the colonic mucus and causes apoptosis by affecting enterocyte permeability (Esposito *et al.,* 2011). CBD improved this damage by inhibiting the apoptotic process and rehabilitating the intestinal barrier integrity via the CB1 receptors (Alhamoruni *et al.,* 2010). Furthermore, CBD may play a role in preventing gut microbiota imbalances, which has potential application for the treatment of GIT disorders and colorectal cancer (Aviello *et al.,* 2012; De Filippis *et al.,* 2011).

1.6.2. Medical relevance of Δ^9 -THC

Numerous studies have reported the therapeutic value of Δ^9 -THC for treating a vast range of ailments which include anorexia, sleep apnoea, nausea, AD, appetite stimulation, Parkinsons disease (PD), dystonia and post-traumatic-stress-disorder (PTSD) (Carley *et al.,* 2018; Khalsa *et al.,* 2022; Todaro, 2012). Saft and colleagues (2018) revealed improved motor scores on the Unified Huntington's Disease Rating Scale (UHDRS) for both chorea (involuntary muscle movements) and dystonia (intermittent muscle contractions which cause twisting or repetitive movements) when patients were treated with Δ^9 -THC. This research also revealed

psychological improvements such as decreased irritability, hypersalivation and apathy (Janavs & Aminoff, 1998; Saft *et al.,* 2018).

Kalaydina and colleagues (2017) showed that Δ^9-THC has the potential to improve gut microbiota stability by preventing the development of a Firmicutes: Bacteroidetes ratio imbalance associated with obesity. This indicates the potential of Δ^9-THC in the treatment or prevention of weight gain and colorectal cancer (Todaro, 2012).

1.6.3. Medical relevance of CBC

Prior to the 1970's Cannabichromene (CBC) was often misidentified as CBD, as both CBD and CBC result in similar retention times in gas chromatography (McPartland & Russo, 2001). CBC has anti-inflammatory properties (Davis & Hatoum, 1983; Turner & ElSohly, 1981) and has demonstrated in mice studies to selectively decreased inflammation associated with gastrointestinal motility by providing analgesic effects (Izzo *et al.,* 2012). This effect is attributed to its interaction with various proteins involved in the control of nociceptive pain (pain caused by body tissue damage). CBC has also been shown to affect CB1 and transient potential-ankyrin type 1 receptors which play roles in pain sensation and hence shows potential as therapeutic agent for the treatment of pain (Maione *et al.,* 2011).

El-Alfy and colleagues (2010) have also reported antidepressant-like actions exerted by CBC, contributing to the characteristic mood elevation from *Cannabis*. Additionally, an antifibrotic property of CBC on human skin sebocytes was reported by Oláh and colleagues (2016) suggesting the aplicability of CBC for the treatment of acne.

1.6.4. Medical relevance of CBG

CBG has been shown to have beneficial effects on neuroinflammatory and neurodegenerative processes through gene expression and modulation (Granja *et al.,* 2012). This cannabinoid influences cell membrane cannabinoid receptor-dependent mechanisms and the expression of genes involved in multiple sclerosis pathophysiology (Borelli *et al.,* 2014). *In vivo* studies have indicated antioxidant and anti-inflammatory activities for CBG, providing potential treatment options for conditions such as colitis, acne, inflammatory bowel disease and psoriasis (Oláh *et al.,* 2016; Wilkinson & Williamson, 2007). Baek and colleagues (1998) investigated the inhibitory effects of CBG and reported its potential for the treatment and prevention of cancers

including colorectal and breast cancer. Furthermore, the antimicrobial properties of CBG have been shown to be superior to those of Δ^9 -THC, CBD and CBC, being active against Gram-positive bacteria and a range of fungi (Nachnani *et al.,* 2021).

1.7. Applications of olivetol and olivetolic acid

OLA and olivetol, a derivative of OLA involved in the biosynthesis of cannabinoids, have a wide range of potential applications across various fields (Kearsey *et al.,* 2019). A study by Lee and colleagues (2023) revealed that OLA has antibacterial activity against *Bacilllus subtilis* at a low minimum inhibitory concentration (MIC). Furthermore, Anderson and colleagues (2022) demonstrated that OLA has modest anticonvulsant activity against hyperthermia-induced seizures in mice models that have Dravet syndrome.

Olivetol has the potential to treat various inflammatory conditions (Gagne *et al.,* 2012; Kearsey *et al.,* 2019). Its anti-inflammatory activity is attributable to its ability to inhibit the production of pro-inflammatory mediators, including cytokines and chemokines (Formukong *et al.,* 1988). Olivetol also modulates the activity of important enzymes involved in inflammation, such as cyclooxygenase (COX) and lipoxygenase (LOX) (Formukong *et al.,* 1988; Hładyszowski *et al.,* 1998). Olivetol further reduces the production of inflammatory enzymes and proteins by suppressing transcription factor nuclear factor-kappa B (NF-κB).

Olivetol and OLA are also rich sources of polyphenols, allowing it to act as an antioxidant (Formukong *et al.,* 1988). The chemical structure of olivetol indicates potent free radical scavenging properties, which offer potential for olivetol to combat oxidative stress by neutralizing reactive oxygen species (ROS) (Garrido *et al.,* 2022). Olivetol enhances the activity of endogenous antioxidant enzymes, which eliminate ROS from the body reducing oxidative stress (Garrido *et al.,* 2022; Hładyszowski *et al.,* 1998). Additionally, this antioxidant activity has also been shown to inhibit lipid peroxidation, thus maintaining the integrity of cell membranes and preserving cellular function (Garrido *et al.,* 2022).

The anti-inflammatory and antioxidant properties of olivetol and OLA indicate their potential applications within a variety of fields, including cosmetic and medicinal fields. (Garrido *et al.,* 2022). Both olivetol and OLA can be used in the cosmetic industry in body creams, lotions and masks as ingredients in nutritional supplements, and as natural preservatives and anti-aging agents (Hładyszowski *et al.,* 1998). In the medical field both have applications for the treatment

of inflammatory conditions, viral and bacterial infections and oxidative stress related ailments (Formukong *et al.,* 1988; Garrido *et al.,* 2022; Hładyszowski *et al.,* 1998).

1.8. Plant genome assembly

Plant genomes provide key insights into their evolution, genetic potential and physiology and can furthermore accelerate the selective breeding capacity of the plant for more desirable phenotypes (Yuan *et al.,* 2017). This has led to a veritable explosion in the sequencing, assembly and analysis of plant genomes, primarily through large-scale projects such as the Earth BioGenome Project (EBP) and the 1000 Plant genome (1K) project (Blaxtera *et al.,* 2022 Coddington *et al.,* 2019; Leebens-Mack *et al.,* 2019; Lewin *et al.,* 2018). Plant genome sequencing escalated during the early 2000s, with an estimated 183 plant reference sequences published by 2016. This number has continued to increase, reaching a total of 7,347 genomes available in the NCBI database for Viridiplantae, including model and non-model species.

The development of Next Generation Sequencing (NGS) in the early 2000s made plant genome sequencing highly accurate and economical (Michael & VanBuren, 2015). The development of NGS platforms including 454, Illumina, and SOLiD, and Ion Torrent between 2005 and 2011 led to a rapid increase in plant genome sequencing (Bentley *et al.,* 2008; Margulies *et al.,* 2005; McKernan *et al.,* 2009; ABI SOLiD launch 2007). These NGS platforms function using different chemistries to reconstruct nucleotide sequences, with most replying on cycle sequencing reactions and imaging (such as Illumina) (Changsheng *et al.,* 2018).

A major shortcoming associated with plant genome assembly using NGS-generated short read data (read lengths of 50-300bp) is the limited ability to cover the full range of the repetitive regions in plant genomes, resulting in fragmented assemblies and collapsed contigs (Changsheng *et al.,* 2018). This leads to assemblies containing significant gaps in the nucleotide sequences and unordered contigs. This shortcoming has recently been addressed through the development of bioinformatic tools that are capable of handling more complex sequence data, together with the advent of Third generation sequencing (TGS). TGS comprises a group of real-time sequencing technologies that directly read single DNA molecules without the need for amplification, generating longer reads and faster sequencing turnaround times compared to NGS (Xiao & Zhou, 2020). Major TGS technologies include Pacific Biosciences (PacBio) (Eid *et al.,* 2009) and Oxford Nanopore Technologies (ONT) (Clarke *et al.,* 2009) sequencing (Changsheng *et al.,* 2018). ONT sequencing allows for real-time analysis of long-

read DNA and RNA through the detection of changes in ionic current as nucleic acids pass through nanopores (Clarke *et al.,* 2009). PacBio Single Molecule Real-Time (SMRT) sequencing is performed on cells using zero-mode waveguides (ZMWs) to observe the incorporation of nucleotides in real time, with read length reaching up to 100kb or more (Eid *et al.,* 2009; Xiao & Zhou, 2020). Long-read sequencing technology can span entire repeat regions, resolve repetitive sequences and decrease errors in plant genome assembly (Changsheng *et al.,* 2018). Due to the complexity and size of plant genomes, long-read TGS is better suited to produce complete complex plant genomes compared to NGS, specifically due to its great potential for application in *de novo* assembly, scaffolding and gap-filling (Xiao & Zhou, 2020). However, currently TGS technologies are costly, precluding their general use for the sequencing of complex plant genomes (Jaio and Schneeberger, 2017). This has resulted in the use of hybrid approaches, where, where both short-read and long-read sequencing is undertaken. The short-reads are used for initial assembly, with lower amounts of long-reads used for gap closure and scaffolding (Bombarely *et al.,* 2016; Yang *et al.,* 2016).

1.9. Genomic studies of *C. sativa*

Genomic resources for *Cannabis* are primarily transcriptome-based, with multiple cultivars having neither a physical nor a genetic map of their genome (Rosenthal, 2007). A genetic map is generated based on recombination frequencies showing the occurrence of two markers being inherited together without indicating the physical distance between these markers. A physical map, however, reflects the actual DNA sequence, including location and physical distance between markers, with distance measured in base pairs (bp) (National Human Genome Research Institute, 2026). Currently, a total of twenty-five genomes for *C. sativa* are available on the National Centre for Biotechnology Information (NCBI) database covering a variety of cultivars and hybrids (Gao *et al.,* 2020; Grassa *et al.,* 2018; Laverty *et al.,* 2019). A large portion of these genome assemblies have significant gaps and inconsistent chromosome nomenclature.

Van Bakel and colleagues (2011) produced the first draft genome assembly and transcriptome for *C. sativa* (Purple Kush cultivar) using Illumina short-read genomic and transcriptomic sequences which were assembled using the SOAPdenovo tool. The draft genome indicated that 70% of the *C. sativa* genome is comprised of repetitive sequence (Pisupati *et al.,* 2018). This study further compared the transcriptomes of Purple Kush and the hemp cultivar Finola using

RNA-sequence transcriptome data (van Bakel *et al.,* 2011). This showed that the cannabinoid biosynthetic pathway genes were more highly expressed in Purple Kush (van Bakel *et al.,* 2011). This was due to the presence and upregulation of Δ^9-THCAS and the artificial selection for THCAS over CBDAS, which led to increased THC levels in drug-type Purple Kush (de Meijer *et al.,* 2009; van Bakel *et al.,* 2011). In contrast, the Finola cultivar have been selected for CBDAS activity for domestication of hemp cultivars for crop production (Morimoto *et al.,* 1998; van Bakel *et al.,* 2011). This study identified a need for *C. sativa* genome assemblies and comparative analysis of *C. sativa* cultivars with altered cannabinoid profiles, for therapeutic applications.

Recently Gao and colleagues (2020) generated a high-quality draft genome for wild-type *C. sativa* varieties using PacBio SMRT sequencing. Analysis of this genome indicated high levels of heterozygosity because of long-term growth in a wild environment, indicating a new resource for genetic information regarding the cannabinoid pathways of *C. sativa* (Gao *et al.,* 2020).

ONT sequencing was used to produce a high-quality chromosome-resolved genome for *C. sativa* CBDRx, a CBD-type cultivar (Grassa *et al.,* 2021). Genomic analysis in this study revealed that CBDAS was expressed as a result of the introgression of this gene (*CBDAS* gene) from hemp. This gene transfer resulted in a shift in the predominant cannabinoid produced from THC to CBD, changing the plant's chemotype from THC-dominant (chemotype I) to CBD-dominant (chemotype III). However, its overall genetic background remained characteristic of drug-type *Cannabis* (Grassa *et al.,* 2021).

1.9.1. Genome repetitiveness and ploidy in *Cannabis*

Notable challenges with plant genome assembly includes the high repetitiveness in the genome due to transposable elements, the polyploid nature of plants and the large size of their genomes, with some reaching up to 22 Gb in size (Neale *et al.,* 2014; Nystedt *et al.,* 2013, Salman-Minkov *et al.,* 2016). Throughout evolutionary history, plant genomes undergo restructuring driven by active transposon amplification (Yuan *et al.,* 2017), with long terminal repeat (LTR) transposons being the most abundant. These elements replicate through reverse transcription followed by reintegration into the genome, typically spanning over 10 kb, and frequently contribute to incomplete and fragmented plant genome assemblies (Changsheng *et al.,* 2018). This repetitiveness and polyploid nature often results in the presence of large gaps in genome

assemblies Research has indicated that ~ 70% of *C. sativa* is composed of repetitive sequences (Pisupati *et al.,* 2018), and as a result *Cannabis* genome assemblies often has large gaps.

Polyploidy refers to the condition of having two or more complete sets of chromosomes. This phenomenon is widespread in plants (Crawford *et al.,* 2021; Parsons *et al.,* 2019) and plays a key role in shaping plant evolution, driving biodiversity, improving crops and helping facilitate domestication (Bagheri & Mansouri, 2015; Mansouri & Bagheri, 2017; Kurtz *et al.,* 2020; Parsons *et al.,* 2019). Polyploidization events occur through two major mechanisms autopolyploidy (a plant with multiple sets of chromosomes derived from genome duplication within a single species) or allopolyploidy (a plant with multiple sets of chromosomes derived from hybridization between two or more different species) (Bagheri & Mansouri, 2015; Mansouri & Bagheri, 2017; Falistocco *et al.,* 2024). Natural ploidy, which refers to the number of complete sets of homologous chromosomes present in a cell or organism, occurs in a plant through whole genome polyploidization events, resulting in plants that display unique phenotypes (Falistocco *et al.,* 2024). Unique phenotypes resulting from polyploidization events often provide the plant with advantages compared to its diploid counterparts, including a morphological or physiological fitness edge, or enhanced adaptability to environmental conditions or stresses (Bajaj *et al.,* 1980; Mansouri & Bagheri, 2017). Beneficial phenotypic traits that have been observed in diploid or triploid plants include larger buds, flowers or leaves and increases in active metabolite concentrations have also been recorded in polyploid plants of medicinal importance including *Artemisia annua*, *Datura stramonium*, *Echinacea purpurea*, *Papaver somniferum, Thymus persicus* (Bajaj *et al.,* 1980; Crawford *et al.,* 2021; Kurtz *et al.,* 2020; Parsons *et al.,* 2019; Sattler *et al.,* 2016).

Cannabis sativa is a diploid species with both male (heterogametic (XY)) and female (homogametic (XX)) plants, with each parent passing down ten chromosomes to their offspring plant (Malabadi *et al.,* 2024; McPartland, 2018; Punja & Holmes, 2020). Diploidy is considered a survival advantage as two sets of chromosomes provides a "genetic" backup if one chromosome is damaged (such as a mutation) (Crawford *et al.,* 2021). However, selective breeding has created varieties that contain three (triploid) or four (tetraploid) sets of chromosomes (Clarke, 1981; Crawford *et al.,* 2021; Kurtz *et al.,* 2020; Parsons *et al.,* 2019). These polyploid *Cannabis* plants have been found to differ in various ways including cell wall composition, biomass production and reduced fertility (Bajaj *et al.,* 1980; Philbrook *et al.,* 2023). *Cannabis* breeders have started utilising various approaches to introduce polyploidy and produce ideal plant varieties including cultivars with higher levels of THC, bigger buds, higher

yields and more flavour (Crawford *et al.,* 2021; Bagheri & Mansouri, 2015; Parsons *et al.,* 2019). These approaches include chemical intervention (colchicine, oryzalin, caffeine, trifuralin, or phosphoric amides) and gaseous intervention (nitrous oxide) (Crawford *et al.,* 2021; Kurtz *et al.,* 2020; Parsons *et al.,* 2019). Thus far, triploid *Cannabis* plants have been found occur naturally in cultivated and wild populations (Philbrook *et al.,* 2023). Triploid plants produce a seedless crop, with bigger yields, larger buds and increased production of phytocannabinoids (Crawford *et al.,* 2021; Bagheri & Mansouri, 2015; Kurtz *et al.,* 2020; Malabadi *et al.,* 2024; Parsons *et al.,* 2019; Sattler *et al.,* 2016). Tetraploid *Cannabis* plants also have unique characteristics which make them desirable in agriculture, including larger leaves, greater trichome density (results in the production of significantly larger volumes of cannabinoids and terpenes) and larger biomass yields overall (Crawford *et al.,* 2021; Kurtz *et al.,* 2020; Malabadi *et al.,* 2024; Sattler *et al.,* 2016). Additionally, tetraploid plants have been recorded to have elevated CBD content (Parsons *et al.,* 2019), however, no tetraploids have been identified outside of a laboratory.

Understanding ploidy and its phenotypic effects in a plant species is key to understanding and exploiting the full potential of this species. However, genome assembly of ploidy plants is extremely challenging, and often results in misassembles (Kyriakidou *et al.,* 2018).

1.10. Single nucleotide variants in *C. sativa*

A single nucleotide variant (SNV) refers to a variation in a DNA sequence that occurs when a single nucleotide in the sequence is altered (Liu *et al.,* 2019). If the SNV is present in at least 1% of the population it is referred to as a single nucleotide polymorphism (SNP).

Single nucleotide variant research on *C. sativa* has indicated that heterozygosity rates range from 0.18-0.26% with drug-type and fibre-type strains being clearly separated by SNVs (van Bakel *et al.,* 2011). The rate of the occurrence of these SNVs between types reaches up to 0.64%. Multiple SNPs in *Cannabis* have been identified in the major cannabinoid genes and used as markers for genetic variation studies in *Cannabis* (Hillig, 2005; Kojoma *et al.,* 2006; Thichak *et al.,* 2011). These SNPs include a set of markers that have been developed for differentiating between *THCA* and *CBDA* synthase genes which successfully separate chemotypes I, II and III (Pacifico *et al.,* 2006). Kojoma and colleagues (2006) identified two distinct variations of the *THCAS* gene, differing by 37 amino acid substitutions, which differentiate drug- and fibre-type *Cannabis*.

1.11. Purpose of this study

Here we have sequenced, assembled and annotated the genome of the Swazi Gold (SG) landrace of *C. sativa*, which originates from the Kingdom of Eswatini (Casano *et al.,* 2010; Crampton, 2015). Swazi Gold has been consistently cultivated over the years with little hybridization from exogenous *Cannabis* cultivars; and as a result, this landrace represents an ancestral, non-hybridized genetic lineage (Casano *et al.,* 2010; Crampton, 2015). It is considered a drug-type plant as it is primarily known and cultivated for its potent psychoactive effects and Δ^9-THC levels (16-18%) (Hartsel *et al.,* 2016). However, little is known about the cannabinoids produced by this landrace, and their potential therapeutic applications. Despite the increasing number of *C. sativa* genome assemblies, very few high-quality genome assemblies are publicly available, and none represent African landraces. As such, this study aimed to sequence, assemble, annotate and compare the genome of this African landrace. The genome of *C. sativa* SG provides a valuable genetic resource to increase our understanding of cannabinoid biosynthetic pathways and explore the evolutionary and functional diversity of *C. sativa* plants with altered chemical profiles. Through the assembly, annotation and comparative analysis this study begins to address this gap and explore the potential the various applications of plants with altered cannabinoid profiles.

1.12. Aim and objectives

This study aimed to assemble and annotate the genome of *Cannabis sativa* landrace Swazi Gold (SG) and perform comparative genome analysis.

The specific objectives of this study included the following:

- Sequencing and assembly of the genome of *C. sativa* SG
- Annotation of the *C. sativa* SG genome to identify key structural and functional genes
- Conducting comparative genomic analyses with *C. sativa* genomes currently available on NCBI

1.13. References

Abrams, D.I., Jay, C.A., Shade, S.B., Vizoso, H., Reda, H., Press, S., Kelly, M.E., Rowbotham, M.C., Petersen, K.L. (2007). '*Cannabi*s in painful HIV-associated sensory neuropathy: a randomized placebo-controlled trial', *Neurology*, 68(7), pp. 515-521. Doi: 10.1212/01.wnl.0000253187.66183.9c

Akbas, F., Gasteyger, C., Sjödin, A., Astrup, A., Larsen, T.M. (2008) 'A critical review of the cannabinoid receptor as a drug target for obesity management'. *Obesity*, 10(1), pp. 58-67. Doi: https://doi.org/10.1111/j.1467-789X.2008.00520.x

Alhamoruni, A., Lee, A.C., Wright, K.L., O'Sullivan, S.E. (2010). 'Pharmacological effects of cannabinoids on the Caco-2 cell culture model of intestinal permeability', *Journal of Pharmacology and Experimental Therapeutics*, 335, pp. 92-102. Doi: 10.1124/jpet.110.168237

Anderson LL, Udoh M, Everett-Morgan D, Heblinski M, McGregor IS, Banister SD, Arnold JC. (2022). 'Olivetolic acid, a cannabinoid precursor in *Cannabis sativa*, but not CBGA methyl ester exhibits a modest anticonvulsant effect in a mouse model of Dravet syndrome', *Journal of Cannabis Research*, 4(1), pp. 1-9. Doi: 10.1186/s42238-021-00113-w.

Andre, C.M., Hausman, J.F., Geurriero, G. (2016). '*Cannabis sativa*: The plant of the thousand and one molecules', *Frontiers in Plant Science*, 7, pp. 1-14. Doi: 10.3389/fpls.2016.00019

Appendino, G., Chianese, G., Taglialatela-Scafati, O. (2011). 'Cannabinoids: Occurrence and medicinal chemistry', *Current Medicinal Chemistry*, 18 (7), pp. 1085-1099. Doi: 10.2174/092986711794940888

Aviello, G., Romano, B., Borrelli, F., Capasso, R., Gallo, L., Piscitelli, F., Marzo, V.D., Izzo, A.A. (2012). 'Chemopreventive effect of the non-psychotropic phytocannabinoid cannabidiol on experimental colon cancer', *Journal of Molecular Med*icine, 90, pp. 925-934. Doi: 10.1007/s00109-011-0856-x

Baek, S.H., Kim, Y.O., Kwag, J.S., Choi, K.E., Jung, W.Y., Han, D.S. (1998). 'Boron trifluoride etherate on silica-A modified Lewis acid reagent (VII). Antitumor activity of cannabigerol against human oral epithelioid carcinoma cells', *Archives of Pharmacal Research*, 21, pp. 353-356. Doi: 10.1007/BF02975301

Bagheri, M., Mansouri, H. (2015). 'Effect of induced polyploidy on some biochemical parameters in *Cannabis sativa* L.', *Applied Biochemistry and Biotechnology*, 175, pp. 2366-2375. Doi: 10.1007/s12010-014-1435-8.

Batchelder, T. (2004). '*Cannabis sativa* and the anthropology of pain." *Townsend Letter for Doctors and Patients*, 247-248, pp. 156. Available at

link.gale.com/apps/doc/A113807045/AONE?u=anon~d0d4fde1&sid=googleScholar&xid=aee6d232. (Accessed 06/04/2024)

Bajaj, Y.P.S., Saini, S.S., Bidani, M. (1980). 'Production of triploid plants from the immature and mature endosperm cultures of rice.', *Theoretical and Applied Genetics*, 58, pp. 17-18. Doi: 10.1007/BF00264662

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Smith, A.J. (2008). 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature,* 456(7218), pp. 53-59. Doi: 10.1038/nature07517

Blaxtera, M., Archibald, J.M., Childers, A.K., Coddington, J.A., Crandall, K.A., Palmag, F.D., Durbina, R., Edwardsi, S.V., Gravesk, J.A.M., Hackettm, Hall, K.J.N., Jarvison, E.D., Johnsonq, R.N., Karlssonr, E.K., Kresst, W.J., Kurakuu, S., Lawniczaka, M.K.N., Lindblad-Tohs, K., Lopezx, J.V., Moranz, N.A., Robinsonaa, G.E., Rydercc, O.A., Shapiroee, B., Soltisff, P.S., Warnowhh, T., Zhangii, G., Lewinkk, H.A. (2022). 'Why sequence all eukaryotes?', *PNAS*, 119, pp. 4. Doi: 10.1073/pnas.2115636118

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C., Bliek, M., Boersma, M., Borghi, L., Bruggmann, R., Bucher, M., D'Agostino, N., Davies, K., Druege, U., Dudareva, N., Egea-Cortines, M., Delledonne, M., Fernandez-Pozo, N., Franken, P., Grandont, L., Heslop-Harrison, J.S., Hintzsche, J., Johns, M., Koes, R., Kuhlemeier, C. (2016). 'Insight into the evolution of the *Solanaceae* from the parental genomes of *Petunia hybrida*', *Nature Plants*, 2 (16074). Doi: 10.1038/nplants.2016.74

Bonini, S.A., Tambaro, S., Kumar, A., Mac-carinelli, G., Memo, M., Mastinu, A. (2018). '*Cannabis sativa:* A comprehensive ethnopharmacological review of a medicinal plant with a long history'. *Journal of Ethnopharmacology*, 227, pp. 300-315. Doi: 10.1016/j.jep.2018.09.004

Borelli, A.C., Freitas, A.E., Silva, R.H., Campos, R.M. (2014). 'Pathophysiology and cell membrane cannabinoid receptor-dependent and receptor-independent mechanisms', *International Journal of Molecular Sciences, 15*(9), Doi: 10.3390/ijms150915653

Bradshaw, R.H.W., Coxon, P., Greig, J.R.A., Hall. A.R. (1981). 'New fossil evidence for the past cultivation and processing of hemp (*Cannabis sativa* L.) in Eastern England', *New Phytologist*, 89, pp. 503-510. Doi: 10.1111/j.1469-8137.1981.tb02331.x

Braich, S., Baillie, R.C., Spangenberg, G.C., Cogan, N.O.I. (2020). 'A new and improved genome sequence of *Cannabis sativa'*, *Gigabyte*, pp. 1-13. Doi: 10.46471/gigabyte.10

Capodice, J.L. & Kaplan, S.A. (2021). 'The endocannabinoid system, *Cannabis*, and cannabidiol: Implications in urology and men's health', *Current Urology*, 15(2), pp. 95-100. Doi: 10.1097/CU9.0000000000000023

Carley, D.W., Prasad, B., Reid, K.J., Malkani, R., Attarian, H., Abbott, S.M., Vern, B., Xie, H., Yuan, C., Zee, P.C. (2018). 'Pharmacotherapy of apnea by cannabimimetic enhancement, the PACE clinical trial: Effects of dronabinol in obstructive sleep apnea', *Sleep*, 41(1), zsx184. Doi: 10.1093/sleep/zsx184

Casano, S., Grassi, G., Martini, V., Michelozzi, M. (2010). 'Variations in terpene profile of different strains *of Cannabis sativa* L.', *ISHS Acta Horticulture*, 925, pp. 115-121. Doi: 10.17660/ActaHortic.2011.925.15

Cervantes, J. (2007). 'Marijuana horticulture: The indoor/outdoor medical grower's bible'. *Marijuana Horticulture: The Indoor/outdoor Medical Grower's Bible*. Available at: https://books.google.co.za/books?hl=en&lr=&id=fERzFsZhdxYC&oi=fnd&pg=PR20&dq=Cervantes,+J.+(2007).+Marijuana+horticulture:+The+Indoor/Outdoor+Medical+Grower%E2%80%99s+bible.+Vancouver,+B.C.:+Van+Patten.&ots=t5Ot8sPJNm&sig=ft4k4ynlzBXIL-yu18ZMGPWde6s&redir_esc=y#v=onepage&q&f=false (Accessed 23/02/2024)

Chagné, D., Batley, J., Edwards, D., Forster, J. W. (2007). 'Single nucleotide polymorphisms genotyping in plants', *Plant Genotyping II: SNP Technology*, 53, pp. 77-94. Doi: 10.1007/978-0-387-36011-9_5

Changsheng, L., Feng, L., Dong, A., Wenqin, W., Ruidong,H. (2018) 'Genome sequencing and assembly by long reads in plants', *Genes*, 9 (1), pp. 6. Doi:10.3390/genes9010006

Chouvy, P.A. (2019). '*Cannabis* cultivation in the world: heritages, trends and challenges', *EchoGeo*, 48, pp. 1-10. Doi: 10.4000/echogeo.17591

Clarke, R.C. & Watson, D.P. (2007). '*Cannabis* and natural *Cannabis* medicines forensic science and medicine: marijuana and the cannabinoids', *Forensic Science and Medicine Humana Press*, pp. 1-15. Doi: 10.1007/978-1-59259-947-9_1

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009). 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nature Nanotechnology,* 4(4), pp. 265-270. Doi: 10.1038/nnano.2009.12

Clarke R.C. & Merlin, M. D. (2013). '*Cannabis*: Evolution and ethnobotany', *University of California Press*. Available at: https://books.google.co.za/books?hl=en&lr=&id=bs4hEAAAQBAJ&oi=fnd&pg=PR11&ots=gLg5h7jJfB&sig=B0T4x-VPrRNKFP6MlufP4YbH73g&redir_esc=y#v=onepage&q&f=false (Accessed 23/02/2024)

Clarke, R.C. (1981). 'The genetics and breeding of *Cannabis*', In *Marijuana Botany: An Advanced Study: The Propagation and Breeding of Distinctive Cannabis,* eds N. Hamel and D. Cross (Berkeley, CA: Ronin Publishing), pp. 27–59. Available at: https://books.google.co.za/books?hl=en&lr=&id=1qyh22v5glgC&oi=fnd&pg=PR16&dq=Clarke,+R.C.+(1981).+%E2%80%98The+genetics+and+breeding+of+Cannabis,,+in+Marijuana+Botany:+An+Advanced+Study:+The+Propagation+and+Breeding+of+Distinctive+Cannabis,+eds+N.+Hamel+and+D.+Cross+(Berkeley,+CA:+Ronin+Publishing),+pp.+27%E2%80%9359.+Doi&ots=IA9W9IMTLQ&sig=5HU5d9oSlUVPPHgNIlAETL3zZUY&redir_esc=y#v=onepage&q&f=false (Accessed 22/06/2025)

Coddington, J., Lewin, H.A., Robinson, G.E., Kress, W.J. (2019). 'The Earth BioGenome Project working group, the Earth BioGenome Project 2020: Starting the clock', *Biodiversity Information Science and Standards*, 3, pp. 1-6. Doi: 10.3897/biss.3.37344

Consroe, P. (1998). 'Brain cannabinoid systems as targets for the therapy of neurological disorders' *Neurobiology of disease*, 5 (6), pp. 534-551. Doi: https://doi.org/10.1006/nbdi.1998.0220

Crampton, H. (2015). 'Dagga. A short history (then, now &just now)', *South Africa: Jacana*. Available at: https://books.google.co.za/books/about/Dagga.html?id=XTjqsgEACAAJ&redir_esc=y (Accessed 16/03/2024)

Crawford, S., Rojas, B.M., Crawford, E., Otten, M., Schoenenberger, T.A., Garfinkel, A.R., Chen, H. (2021). 'Characteristics of the diploid, triploid, and tetraploid versions of acannabigerol-dominant F1 hybrid industrial hemp cultivar, *Cannabis sativa* 'stem cell CBG'', *Genes (Basel),* 12(6), pp. 923. Doi: 10.3390/genes12060923.

Davis, W.M. & Hatoum, N.S. (1983). 'Neurobehavioral actions of cannabichromene and interactions with delta-9-tetrahydrocannabinol', *General Pharmacology*, 14(2), pp. 247-252. Doi: 10.1016/0306-3623(83)90004-6

De Filippis, D., Esposite, G., Cirillo, C., Cipriano, M., Winter, B.Y.D., Scuderi, C., Sarnelli, G., Cuoma, R., Steardo, L., Man, J.G., Iuvone, T. (2011). 'Cannabidiol reduces intestinal inflammation through the control of neuroimmune axis', *PloS One*, 6. Doi: 10.1371/journal.pone.0028159

De Meijer, E. & Hammond, K. (2005). 'The inheritance of chemical phenotype in *Cannabis sativa* L. (II): Cannabigerol predominant plants', *Euphytica,* 145. pp. 189-198. Doi: 10.1007/s10681-005-1164-8

De Meijer, E.P.M., Hammond, K., Micheler, M. (2009). 'The inheritance of chemical phenotype in *Cannabis sativa* L. (III): variation in cannabichromene proportion', *Euphytica*, 165, pp. 293-311. Doi: 10.1007/s10681-008-9787-1

De Meijer, E., Hammond, K., Sutton, A. (2009). 'The inheritance of chemical phenotype in *Cannabis sativa* L. (IV): Cannabinoid-free plants', *Euphytica*, 168, pp. 95-112. Doi: 10.1007/s10681-009-9894-7

De Meijer, E. & Hammond, K. (2016). 'The inheritance of chemical phenotype in *Cannabis sativa* L. (V): Regulation of the propyl-/pentyl cannabinoid ratio, completion of a genetic model', *Euphytica*, 210, pp. 291-307. Doi: 10.1007/s10681-016-1721-3

Deamer, D., Akeson, M., Branton, D. (2016). 'Three decades of nanopore sequencing', *Nature Biotechnology*, 34, pp. 518-524. Doi: 10.1038/nbt.3423

Dewick, P.M. (2009). 'Medicinal natural products: A biosynthetic approach', *3rd ed. John Wiley & Sons: Chichester*. Available at: https://books.google.co.za/books?hl=en&lr=&id=A4zptjOJfKQC&oi=fnd&pg=PA1&dq=Dewick,+P.M.+Medicinal+Natural+Products:+A+Biosynthetic+Approach,+3rd+ed.%3B+John+Wiley+%26+Sons:+Chichester,+UK,+2009&ots=wjdpF-NEYh&sig=wDOcPV0KU6LIyKpOpUoagFQJv8c&redir_esc=y#v=onepage&q&f=false (Accessed 20/03/2024)

Duggan, P.J. (2021). 'The chemistry of *Cannabis* and cannabinoids', *Australian Journal of Chemistry*, 74, pp. 369-387. Doi: 10.1071/CH21006

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dala, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., King, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers., K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S. (2009). 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323 (5910), pp. 133-138. Doi: 10.1126/science.1162986

El-Alfy, A.T., Ivey, K., Robinson, K., Ahmed, S., Radwan, M., Slade, D., Khan, I., ElSohly, M., Ross, S. (2010). 'Antidepressant-like effect of delta9- tetrahydrocannabinol and other cannabinoids isolated from *Cannabis sativa* L', *Pharmacology Biochemistry and Behaviour*, 95(4), pp. 434-442. Doi: 10.1016/j.pbb.2010.03.004

ElSohly, M.A., Radwan, M.M., Gul, W., Chandra, S., Galal, A. (2017). 'Phytochemistry of *Cannabis sativa* L.', *Progression in the Chemistry of Organic Natural Prododucts*; 103, pp. 1-36. Doi: 10.1007/978-3-319-45541-9_1.

Esposito G, Scuderi C, Valenza M, Giuseppina Ines Togna, Valentina Latina, Daniele De Filippis, Mariateresa Cipriano, Maria Rosaria Carratù,Teresa Iuvone, Luca Steardo. (2011). 'Cannabidiol

reduces Aβ-induced neuroinflammation and promotes hippocampal neurogenesis through PPARγ involvement', *PloS On*e, 6, e28668-e28668. Doi: 10.1371/journal.pone.0028668

Falistocco, E, Prieto, P., Ceccarelli, M., Farooq, M.A. (2024). 'Editorial: Advances in the study of polyploid evolution in wild populations', *Frontiers in Plant Science*, 14, pp. 6-9. Doi: 10.3389/fpls.2023.1335981.

Farag, S. & Kayser,O. (2017). 'The *Cannabis* plant: botanical aspects', *Handbook of Cannabis and Related Pathologies: Biology- Pharmacology*, pp. 3-12. Doi: 10.1016/B978-0-12-800756-3.00001-6

Formukong, E.A., Evans, A.T., Evans, F.J. (1988). 'Analgesic and anti-inflammatory activity of constituents of *Cannabis sativa* L.', *Inflammation*, 12(4), pp. 361-371. Doi: 10.1007/BF00915771

Freeman, M., Petrilli, K., Lees, R., Hindocha, C., Mokrysz, C., Curran, V., Saunders, R., Freeman, T.P. (2019). 'How does cannabidiol (CBD) influence the acute effects of delta-9-tetrahydrocannabinol (THC) in humans? A systematic review', *Neuroscience & Biobehavioral Reviews*, 107, pp. 696-712. Doi: 10.1016/j.neubiorev.2019.09.036

Fidan, M., Süzerer, V., Onay, A. (2023). '*Cannabis sativa* L.*:* Origin, distribution, taxonomy and biology'*, Türk Bilimsel Derlemeler Dergisi*, 16(1), pp. 10-28. Available at: https://derleme.gen.tr/index.php/derleme/article/view/454 (Accessed 18/02/2024)

Gagne, S.J., Stout, J.M., Liu, E., Page, J.E. (2012). 'Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides', *Proceedings of the National Academy of Sciences USA*, 109(31), pp. 12811-12816. Doi: 10.1073/pnas.1200330109

Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., Yu, Y., Yang, W., Zhang, Y. (2020). 'A high-quality reference genome of wild *Cannabis sativa'*, *Horticulture Research*, 7(1), pp.73. Doi:10.1038/s41438-020-0295-3.

Garrido, J., Rico, S., Corral, C., Sánchez, C., Vidal, N., Martínez-Quesada, J.J., Ferreiro-Vera, C. (2022). 'Exogenous application of stress-related signaling molecules affect growth and cannabinoid accumulation in medical *cannabis* (*Cannabis sativa* L.)', *Frontiers in Plant Science*,13, pp.15. Doi: 10.3389/fpls.2022.1082554

Gigli, S., Seguella, L., Pesce, M., Bruzzese, E., D'Alessandro, A., Cuoma, R., Steardo, L., Sarnelli, G., Esposito, G. (2017). 'Cannabidiol restores intestinal barrier dysfunction and inhibits the apoptotic process induced by *Clostridium difficile* toxin A in Caco-2 cells', *United European Gastroenterology Journal*, 5 (8), pp. 1108-1115. Doi: 10.1177/205064061769862

Granja, A.G., Carrillo-Salinas, F., Pagani, A., Gómez-Cañas, M., Negri, R., Navarrete, C., Mecha, M., Mestre, L., Fiebich, B.L., Cantarero, I., Calzado, M.A., Bellido, M.L., Fernandez-Ruiz, J.,

Appendino, G., Guaza, C., Muñoz, E. (2012). 'A cannabigerol quinone alleviates neuroinflammation in a chronic model of multiple sclerosis', *Journal of Neuroimmune Pharmacology,* 7, pp. 1002-1016. Doi: 10.1007/s11481-012-9399-3

Grassa, C.J., Wegner, J.P., Dabney, C., Poplawski, S.G., Motley, S.T., Michael, T.P., Schwartz, C.J., Weiblen, G.D. (2018). 'A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content', *bioRxiv*, Doi:10.1101/458083.

Grivennikov, S.I., Greten, F.R., Karin, M. (2010).'Immunity, inflammation, and cancer," *Cell*, 140 (6), pp. 883-899, 2010. Doi: 10.1016/j.cell.2010.01.025

Grotenhermen, F. (2003). 'Clinical Pharmacodynamics of cannabinoids', *Journal of Cannabis Therapeutics*, 4(1), pp. 29-78. Doi: 10.1300/J175v04n01_03

Grotenhermen, F. (2005). 'Cannabinoids' *Current Drug Targets-CNS & Neurological Disorders,* 4(5), pp. 507-530(24). Doi: 10.2174/156800705774322111

Gulck, T. & Moller, B.L. (2020). 'Phytocannabinoids: origins and biosynthesis.' *Trends in Plant Science*, 25, pp. 985-1004. Doi: 10.1016/j.tplants.2020.05.005

Hanuš, L.O., Meyer, S.M., Muñoz, E., Taglialatela-Scafati, O., Gippendino, G. (2016). 'Phytocannabinoids: a unified critical inventory', *Natural Product Reports*, 33, pp. 1357-1392. Doi: 10.1039/C6NP00074F

Happyana, N., Agnolet, S., Muntendam, R., Van Dam, A., Schneider, B., O. Kayser, O. (2013). 'Analysis of cannabinoids in laser-microdissected trichomes of medicinal *Cannabis sativa* using LCMS and cryogenic NMR', *Phytochemistry*, 87, pp. 51-59. Doi: 10.1016/j.phytochem.2012.11.001

Happyana, N. & Kayser, O. (2016). 'Monitoring metabolite profiles of *Cannabis sativa* L. trichomes during flowering period using H NMR-based metabolomics and real-time PCR', *Planta Medica*., 82, pp. 1217-1223. Doi: 10.1055/s0042-108058

Hartsel, J.A., Eades, J., Hickory, B., Makriyannis, A. (2016). '*Cannabis sativa* and hemp', *Nutraceuticals: Efficacy, Safety and Toxicity*, pp. 735-754. Doi: 10.1016/B978-0-12-802147-7.00053-X

Hazenkamp, A., Jouirou, B., Mouth, L., Sabatier, J.M. (2010). 'Chemistry of *Cannabis*. Comprehensive', *Natural Products II (Elsevier)*, pp. 1033-1084. Doi: 10.1016/b978-008045382-8.00091-5

He, B., Zhao, S., Chen, Y., Cao, O., Wei, C., Cheng, X., Zhang, Y. (2015). 'Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms' *BMC Genomics*, 16, pp. 65. Doi: 10.1186/s12864-014-1192-7

Hesami, M., Pepe, M., Alizadeh, M., Rakei, A., Baiton, A., Jones, A.M.P. (2020). 'Recent advances in *Cannabis* biotechnology', *Industrial Crops and Products*, 158, pp. 113026. Doi: 10.1016/ j.indcrop.2020.113026

Hesami, M., Pepe, M., Jones, A.M.P. (2023). 'Morphological characterization of *Cannabis sativa* L. throughout its complete life cycle', *Plants*, 12, p. 20. Doi: h10.3390/plants12203646

Hillig, K.W. (2005). 'Genetic evidence for speciation in *Cannabis (Cannabaceae*)', *Genetic Resources and Crop Evolution*, 52, pp. 161–180. Doi: 10.1007/s10722-003-4452-y

Hillig, K. & Mahlberg, P. (2005). 'A chemotaxonomic analysis of cannabinoid variation in *Cannabis* (*Cannabaceae*)', *American Journal of Botany*, 91(6), pp. 966-975. Doi: 10.3732/ajb.91.6.966

Hładyszowski J, Zubik L, Kozubek A. (1998). 'Quantum mechanical and experimental oxidation studies of pentadecylresorcinol, olivetol, orcinol and resorcinol', *Free Radical Research*, 28(4), pp. 359-368. Doi: 10.3109/10715769809070804

Huchelmann, A., Boutry, M., Hachez, C. (2017). 'Plant glandular trichomes: natural cell factories of high biotechnological interest', *Plant Physiology*, 175, pp. 6-22. Doi:10.1104/pp.17.00727

Hurgobin, B., Tamiru-Oli, M., Welling, M.T., Doblin, M.S., Whelan, J., Lewsey, M.G. (2020). 'Recent advances in *Cannabis sativa* genomics research', *New Phytologist Foundation*. 230(1), pp. 73-89. Doi: 10.1111/nph.17140

International Rice Genome Sequencing Project: The map-based sequence of the rice genome. (2005). *Nature*, 436 (7052), pp. 793-800. Doi: 10.1038/nature03895

Izzo, A.A., Capasso, R., Aviello, G., Borelli, F., Romano, B., Piscitelli, F., Gallo, L., Capasso, F., Orlando, P. (2012). 'Inhibitory effect of cannabichromene, a major non-psychotropic cannabinoid extracted from *Cannabis sativa*, on inflammation-induced hypermotility in mice', *British Journal of Pharmacology*, 166, pp. 1444–1460. Doi: 10.1111/j.1476-5381.2012.01879.x

Janavs, J.L. & Aminoff, M.J. (1998). 'Dystonia and chorea in acquired systemic disorders', *Journal of Neurology, Neurosurgery & Psychiatry,* 65, pp**.** 436-445. Doi: 10.1136/jnnp.65.4.436

Jarvis, E.D. (2016). 'Perspectives from the Avian Phylogenomics Project: Questions that can be answered with sequencing all genomes of a vertebrate class', *Annual Review of Animal Biosciences*, 4, pp. 45-59. Doi: 10.1146/annurev-animal-021815-111216

Jiao, W. & Schneeberger, K. (2017). 'The impact of third generation genomic technologies on plant genome assembly,' *Current Opinion in Plant Biology* 2017, 36, pp. 64–70. Doi: 10.1016/j.pbi.2017.02.002

Kalaydina, R.V., Bessi Qorri, B., Szewczuk, M.R. (2017). 'Preventing negative shifts in gut microbiota with *Cannabis* therapy: implications for colorectal cancer', *Advanced Research in Gastroenterology & Hepatology*, 7(3), pp. 1-5. Doi: 10.19080/ARGH.2017.07.555712

Kanabus, J., Bryła, M., Roszko, M., Modrzewska, M., Pierzgalski, A., (2021). 'Cannabinoids - Characteristics and potential for use in food production', *Molecules*, 26(21). Doi: 10.3390/molecules26216723

Kearsey, L.J., Prandi, N., Karuppiah, V., Yan, C., Leys, D., Toogood, H., Takano, E., Scrutton, N.S. (2019). 'Structure of the *Cannabis sativa* olivetol-producing enzyme reveals cyclization plasticity in type III polyketide synthases', *The FEBS Journal*. Doi: 10.1111/febs.15089

Kim, A.L., Yun, Y.J., Choi, H.W., Hong, C., Shim, H.J., Lee, J.H., Kim, Y. (2022). 'Profiling cannabinoid contents and expression levels of corresponding biosynthetic genes in commercial *Cannabis* (*Cannabis sativa* L.) cultivars', *Plants*, 11(22), pp. 3088. Doi: 10.3390/plants11223088

Khalsa, J.H., Bunt, G., Blum, K., Maggirwar, S.B., Galanter, M., Potenza, M.N. (2022). 'Review: Cannabinoids as medicinals', *Current Addiction Reports*, 9, pp. 630–646. Doi: 10.1007/s40429-022-00438-3

Kojoma, M., Seki, H., Yoshida, S., Muranaka, T. (2006). 'DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L.', *Forensic Science International*, 159(2-3), pp. 132–140. Doi: 10.1016/j.forsciint.2005.07.005

Kurtz, L.E., Brand, M.H., Lubell-Brand, J.D. (2020). 'Production of tetraploid and triploid hemp', *HortScience*, 55, pp. 1703-1707. Doi: 10.21273/HORTSCI15303-20

Kyriakidou, M., Tai, H.H., Anglin, N.L., Ellis, D., Strömvik, M.V. (2018). 'Current strategies of polyploid plant genome sequence assembly', *Frontiers in Plant Science*, 9(1660). Doi: 10.3389/fpls.2018.01660

Laverty, K.U., Stout, J.M., Sullivan, M.J., Shah, H., Gill, N., Holbrook, L., Deikus, G., Sebra, R., Hughes, T.R., Page, J.E., Van Kabel, H. (2019). 'A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci', *Genome Research*, 29, pp. 146-156. Doi: 10.1101/gr.242594.118.

Lee, YE., Kodama, T., Morita, H. (2023). 'Novel insights into the antibacterial activities of cannabinoid biosynthetic intermediate, olivetolic acid, and its alkyl-chain derivatives', *Journal of Natural Medicines*, 77, pp. 298-305. Doi: 10.1007/s11418-022-01672-9

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis,

D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., DePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Diop, S.I., Elliott, A., Ene, C.O., Eom, A.H., Feng, T., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gitzendanner, M.A., Graham, S.W., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Juniper, B., Kates, H.R., Kessler, M., Kidner, C.A., Kohchi, T., Kovar, L., Krause, K., Krupinska, K., Kuo, L.-Y., Kuo, Y.-M., Labadie, K., Lalonde, M., Lam, V.K.Y., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Li, J., Li, Q., Li, Z., Liu, B., Liu, C., Liu, J., Liu, X., Llorens, C., Luan, S., Matasci, N., McAbee, J.M., McKain, M.R., McLellan, T., McNeal, J.R., McPherson, M.A., McSteen, P., Mendez, L., Mikulic, P., Ming, R., Mower, J.P., Murch, S.J., Nelson, D.R., Nevill, P.G., Newell, C.A., Newton, A.E., Nge, F.J., Nguyen, T., Noutsos, C., Novikova, P.Y., Oldenkott, B., O'Neill, C.M., O'Reilly, J.E., Otto, S.P., Owens, G.L., Petersen, G., Pires, J.C., Poncet, V., Proctor, V., Radhakrishnan, D., Ramesh, S., Rest, J.S., Rexroad, C., Rieseberg, L.H., Rothfels, C.J., Sage, R.F., Sakai, A.K., Salmon, A., Sander, R., Sanguinet, K.A., Schilling, E.E., Schmutz, J., Schulz, M.H., Schwartz, J.A., Scott, A.D., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.R., Smithson, A., Stewart, C.N. Jr., Stinchcombe, J.R., Surek, B., Thomas, P., Villarreal, J.C., Waller, D.M., Warnow, T., Weng, J.-K., Wicke, S., Wu, X., Wullschleger, S.D., Yang, Y., Zhu, D., Zhuang, C., Zinkgraf, M., Wong, G.K.-S., Matasci, N. (2019). 'One thousand plant transcriptomes and the phylogenomics of green plants', *Nature*, 574(7780), pp. 679-685. Doi: 10.1038/s41586-019-1693-2

Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R., Edwards, S.V., Forest, F., Gilbert, M.T., Goldstein, M.M., Grigoriev, I.V., Hackett, K.J., Haussler, D., Jarvis, E.D., Johnson, W.E., Patrinos, A., Richards, S., Castilla-Rubio, J.C., van Sluys, M., Soltis, P.S., Xu, X., Yang, H., Zhang, G. (2018). 'Earth BioGenome Project: Sequencing life for the future of life', *Proceedings of the National Academy of Sciences of the United States of America,* 115, pp. 4325-4333. Doi: 10.1073/pnas.1720115115

Liang, Y.C., Huang, C.C., Hsu, K.S. (2004). 'Therapeutic potential of cannabinoids in trigeminal neuralgia', *Current Drug Targets-CNS & Neurology Disorders*, 3, pp. 507- 514. Doi: 10.2174/1568007043336833

Liu, F., Zhang, Y., Zhang, L., Fang, Q., Gao, R., Zhang, Z., (2019). 'Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data', *Genome Biology*, 20 (242), article 24. Doi: 10.1186/s13059-019-1863-4

Lofgren, E.T., Cole, S.R., Weber, D.J., Anderson, D.J., Moehring, R.W. (2014). 'Hospital-acquired *Clostridium difficile* infections: Estimating all-cause mortality and length of stay', *Epidemiology*, 25, pp. 570-575. Doi: 10.1097/EDE.0000000000000119

Logacheva, M.D., Kasianov, A.S., Vinogradov, D.V., Samigullin, T.H., Gelfand, M.S., Makeev, V.J., Penin, A.A. (2011), '*De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*)', *BMC Genomics*, 12 (30). Doi: 10.1186/1471-2164-12-30

Long, T., Long, T., Wagner, M., Demske, D., Leipe, C., Tarasov, P.E. (2017). '*Cannabis* in Eurasia: origin of human use and Bronze Age trans-continental connections', *Vegetation History and Archaeobotany*, 26, pp. 245-258. Doi: 10.1007/s00334-016-0579-6

Lowe, H.; Toyang, N.; Steele, B.; Bryant, J.; Ngwa, W. (2021). 'The endocannabinoid system: A potential target for the treatment of various diseases', *International Journal of Molecular Sciences*, 22, pp. 9472. Doi: 10.3390/ijms22179472

Luo, X., Luo, X., Reiter, M.A., d'Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., Lin, W., Lee, H., Yu, C., Shin, J., Deng, K., Benites, V.T., Wang, G., Baidoo, E.E.K., Chen, Y., Dev, I., Petzold, C.J., Keasling, J.D. (2019). 'Complete biosynthesis of cannabinoids and their unnatural analogues in yeast', *Nature*, 567, pp. 123-126. Doi: doi.org/10.1038/s41586-019-0978-9

Lynch, R.C., Vergara, D., Tittes, S., White, K., Schwartz, C.J., Gibbs, M.J., Ruthenburg, T.C., DeCesare, K., Land, D.P., Kane, N.C. (2016). 'Genomic and chemical diversity in *Cannabis*'. *Critical reviews in plant sciences,* 35 (5-6), pp. 349-363. Doi: 10.1080/07352689.2016.1265363

Malabadi, R.B., Kolkar, K.P., Chalannavar, R.K., Baijnath, H. (2024). '*Cannabis sativa*: Monoecious species and Hermaphroditism: Feminized seed production- A breeding effort', *World Journal of Biology Pharmacy and Health Sciences*, 20(03), pp. 169-183. Doi: 10.30574/wjbphs.2024.20.3.0968

Maione, S., Piscitelli, F., Gatta, L., Vita, D., Pettrocellis, L.D., Palazzo, E., de Novellis, V., Marzo, V.D. (2011). 'Non-psychoactive cannabinoids modulate the descending pathway of antinociception in anaesthetized rats through several mechanisms of action', *British Journal of Pharmacology*, 162, pp. 584-596. Doi: 10.1111/j.1476-5381.2010.01063.x

Mansouri H, Bagheri M. (2017). 'The induction of polyploidy and its effect on *Cannabis sativa* L', In *Cannabis sativa* L. - Botany and biotechnology, eds S. Chandra, H. Lata, and M. A. Elsohly (New York, NY: Springer International Publishing). Doi: 10.1007/978-3-319-54564-6_17

Manzanares, J., Julian, M.D., Carrascosa, A. (2006) 'Role of the cannabinoid system in pain control and therapeutic implications for the management of acute and chronic pain episodes', *Current Neuropharmacology*, 4(19), pp. 239-257. Doi: https://doi.org/10.2174/157015906778019527

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M.S., Chen, Y., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J., Rothberg, J.M. (2005). 'Genome sequencing in microfabricated high-density picolitre reactors', *Nature, 437*(7057), pp. 376-380. Doi: 10.1038/nature03959

Marks, M.D., Tian, L., Wenger, J.P., Omburo, S.N., Soto-Fuentes, W., Gang, D.R., Weiblen, G.D., Dixon, R.A. (2009). 'Identification of candidate genes affecting delta(9)-tetrahydrocannabinol biosynthesis in *Cannabis sativa'*, *Journal of Experimental Bot*ony, 60, pp. 3715-3726. Doi: 10.1093/jxb/erp210

McCoy, R.C., Taylor, R.W., Blauwkamp, T.A, Kelley, J.L., Kertesz, M., Pushkarev, D., Petrov, D.A., Fiston-Lavier, A.S. (2014). 'Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements', *PLoS One*, Doi: 10.1371/journal.pone.0106689

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C.L., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M., Blanchard, A.P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding', *Genome Research*, 19(9), pp. 1527-1541. Doi: 10.1101/gr.091868.109.

McPartland, J. M. (2018). '*Cannabi*s systematics at the levels of family, genus, and species', *Cannabis Cannabinoid Research*. 3 (1), pp. 23-212. Doi: 10.1089/can.2018.0039

McPartland, J. & Russo, E.B. (2001). '*Cannabis* and *Cannabis* extracts: Greater than the sum of their parts', *Journal of Cannabis Therapeutics*, 1(3-4), pp. 103-132. Doi: 10.1300/J175v01n03_08

Michael, T.P. & VanBuren, R. (2015). 'Progress, challenges and the future of crop genomes', *Current Opinion in Plant Biology*, 24, pp. 71-81. Doi: 10.1016/j.pbi.2015.02.002

Morimoto, S., Komatsu, K., Taura, F., Shoyama, Y. (1998). 'Purification and characterization of cannabichromenic acid synthase from *Cannabis sativa*', *Phytochemistry,* 49, pp. 1525-1529. Doi: 10.1016/S0031-9422(98)00278-7

Nachnani, R., Raup-Konsavage, W.M., Vrana, K.E. (2021). 'The Pharmacological Case for Cannabigerol', *The Journal of Pharmacology and Experimental Therapeutics*, 376 (2), pp. 204-212. Doi: 10.1124/jpet.120.000340

Namdar, D. & Koltai, H. (2018). "Medical *Cannabis* for the treatment of inflammation," *Natural Product Communications*, 13, pp. 249-254. Doi: 10.1177/1934578X180130030

National Human Genome Research Institute. (2026). "Mapping. Genetics-Glossary" Available at: https://www.genome.gov/genetics-glossary/Mapping (Accessed 09/02/2026)

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., Martínez-García, P.J., Vasquez-Gross, H.A., Lin, B.Y., Zieve, J.J., Dougherty, W.M., Fuentes-Soriano, S., Wu, L., Gilbert, D., Marçais, G., Roberts, M., Holt, C., Yandell, M., Davis, J.M., Smith, K.E., Langley, C.H. (2014). 'Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies', *Genome Biology*, 15, R59. Doi: 10.1186/gb-2014-15-3-r59

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Jansson, S. (2013). 'The Norway spruce genome sequence and conifer genome evolution', *Nature*, 497, pp. 579-584. Doi: 10.1038/nature12211

Oláh, A., Markovics, A., Szabó-Papp, J., Szabó, P.T., Stott, C., Zouboulis, C.C., Biró, T. (2016). 'Differential effectiveness of selected non-psychotropic phytocannabinoids on human sebocyte functions implicates their introduction in dry/seborrhoeic skin and acne treatment', *Experimental Dermatology,* 25(9), pp. 701-707. Doi: 10.1111/exd.13042

Pacifico, D., Miselli, F., Micheler, M., Carboni, A., Ranalli, P., Mandolino, G. (2006). 'Genetics and marker-assisted selection of the chemotype in *Cannabis sativa* L.', *Molecular Breeding*, 17, pp. 257-268. Doi: 10.1007/s11032-005-5681-x

Parsons, J.L., Martin, S.L., James, T., Golenia, G., Boudko, E.A., Hepworth, S.R. (2019). 'Polyploidization for the genetic improvement of *Cannabis sativa'*, Frontiers in Plant Science, 10, pp. 476. Doi: 10.3389/fpls.2019.00476.

Pellati, F., Borgonetti, V., Brighenti, V., Biagi, M., Benvenuti, S., Corsi, L. (2018). '*Cannabis sativa* L. and non-psychoactive cannabinoids: Their chemistry and role against oxidative stress, inflammation, and cancer', *Hindawi BioMed Research International Volume*, vol. 2018, pp. 15. Doi: 10.1155/2018/1691428

Philbrook, R., Jafari, M., Gerstenberg, S., Say, K.L., Warren, J., Jones, A.M.P. (2023). 'Naturally occurring triploidy in *Cannabis*', *Plants*, 12, pp. 3927. Doi: 10.3390/ plants12233927.

Piluzza, G., Delogu, G., Cabras, A., S. Bullitta. (2013). 'Differentiation between fiber and drug types of hemp (*Cannabis sativa* L.) from a collection of wild and domesticated accessions', *Genetic Resources and Crop Evolution*, 60, pp. 2331-2342. Doi: https://doi.org/10.1007/s10722-013-0001-5

Pisupati, R., Vergara, D., Kane, N.C. (2018). 'Diversity and evolution of the repetitive genomic content in *Cannabis sativa'*, BMC Genomics 19, 156. Doi:10.1186/s12864-018-4494-3

Punja, Z.K., Holmes, J.E. (2020). 'Hermaphroditism in marijuana (*Cannabis sativa* L.) inflorescences – Impact on floral morphology, seed formation, progeny sex ratios, and genetic variation', *Frontiers in Plant Science*, 11, pp. 718. Doi: 10.3389/fpls.2020.00718.

Radwan, M.M., Chandra, S., Gul, S., ElSohly, M.S. (2021). 'Cannabinoids, phenolics, terpenes and alkaloids of *Cannabis*', *Molecules*, 26(9), pp. 2774. Doi: 10.3390/molecules26092774

Raharjo, T.J., Chang, W., Choi, Y.H., Peltenburg-Looman, A.M.P., (2003). 'Olivetol as product of a polyketide synthase in *Cannabis sativa* L', *Plant Science*, 166 (2), pp. 381-385. Doi: 10.1016/j.plantsci.2003.09.027

Ranalli, P. (1999). 'Advances in hemp research', *CRC Press*, pp. 272. Doi: 10.1201/9781498705820

Rosenthal, E. (2007). 'More marijuana varieties from the world's great seed breeders', *The Big Book of Buds*, 3. Available at: https://books.google.co.za/books?hl=en&lr=&id=jc47AAAAQBAJ&oi=fnd&pg=PT7&dq=Rosenthal+E:+The+Big+Book+of+Buds,+Volume+3:+More+Marijuana+Varieties+from+the+World%E2%80%99s+Great+Seed+Breeders+2007.&ots=i3zcZlkJTF&sig=NGvumYNopy6f-7SANti4E16UW4s&redir_esc=y#v=onepage&q&f=false (Accessed 20/03/2024)

Rotherham, D. & Harbison, S.A. (2010). 'Differentiation of drug and non-drug *Cannabis* using a single nucleotide polymorphism (SNP) assay', *Forensic Science International*, 207, pp. 193-197. Doi: 10.1016/j.forsciint.2010.10.006

Russo, EB. (2007). 'History of *Cannabis* and its preparations in saga, science and sobriquet'. *Chemistry and Biodiversity*, 4 (8), pp. 1614-1648. Doi: 10.1002/cbdv.200790144

Saft, C., vo Hein, S.M., Lucke, T., Thiels, C., Peball, M., Diamshidian, A., Heim, B., Seppi, K. (2018). 'Cannabinoids for treatment of dystonia in Huntington's Disease', *Journal of Huntington's Disease*, 7, pp. 167-173. Available at: https://content.iospress.com/articles/journal-of-huntingtons-disease/jhd170283 (Accessed 15/03/2024)

Salentijn, E. M. J., Zhang, Q., Amaducci, A., Yan, M., Trindade, L.M. (2015). 'New developments in fiber hemp (*Cannabis sativa* L.) breeding', *Industrial Crops and Products*, 68, pp. 32-41. Doi: 10.1016/j.indcrop.2014.08.011

Salman-Minkov, A., Sabath, N., Mayrose, I. (2016). 'Whole-genome duplication as a key factor in crop domestication', *Nature Plants*, 2, 16115. Doi: 10.1038/nplants.2016.115

Sattler, M.C., Carvalho, C.R., Clarindo, W.R. (2016). 'The polyploidy and its key role in plant breeding', *Planta*, 243, pp. 281-296. Doi: 10.1007/s00425-015- 2450-x.

Sawler, J., Stout, J.M., Gardner, K.M., Hudson, D., Vidmar, J., Butler, L., Page, J.E., Myles, S. (2015). 'The genetic structure of marijuana and hemp.' *PLoS One*, 10(8). Doi: 10.1371/journal.pone.0133292

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A. *et al.* (2009). 'The B73 maize genome: complexity, diversity, and dynamics', *Science*, 326 (5956), pp. 1112-1115. Doi: 10.1126/science.1178534

Shevyrin, V.A., & Morzherin, Y.Y. (2015). 'Cannabinoids: structures, effects, and classification', *Russian Chemical Bulletin*, 64, pp. 1249-1266. Doi: 10.1007/s11172-015-1008-1

Sirikantaramas, S., Morimoto, S., Shoyama, Y., Ishikawa, Y., Wada, Y., Shoyama, Y., Taura, F. (2004). 'The gene controlling marijuana psychoactivity: Molecular cloning and heterologous expression of 1-tetrahydrocannabinolic acid synthase from *Cannabis sativa* L.', *Journal of Biological Chemistry*, 279 (38), pp. 39767–39774. Doi: 10.1074/jbc.M403693200

Small, E. & Cronquist, A. (1976). 'A practical and natural taxonomy for cannabis', *Taxonomy*, 25, pp. 405-435. Doi: 10.2307/1220524

Spitzer-Rimon, B., Shafran-Tomer, H., Gottlieb, G.H., Doron-Faigenboim, A., Zemach, H., Kamenetsky-Goldstein, R., Flaishman, M. (2022). 'Non-photoperiodic transition of female *cannabis* seedlings from juvenile to adult reproductive stage', *Plant Reproduction*, 35, pp. 265-277. Doi: 10.1007/s00497-022-00449-0

Steel, L., Welling, M., Ristevski, N., Johnson, K., Gendall, A. (2023). 'Comparative genomics of flowering behaviour in *Cannabis sativa*', *Frontiers in Plant Science*, 14, Doi: 10.3389/fpls.2023.1227898

Stout, J.M., Boubakir, Z., Ambrose, S.J., Purves, R.W., Page, J.E. (2012). 'The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl activating enzyme in *Cannabis sativa* trichomes', *The Plant Journal*, 71, pp. 353-365. Doi: 10.1111/j.1365-313X.2012.04949.x

Tahir, M.N., Raz, F.S., Rondeau-Gagné, S., Trant, J.F. (2021). 'The biosynthesis of the cannabinoids', *Journal of Cannabis Res*earch, 3, article 7. Doi: 10.1186/s42238-021-00062-4

Tan, Z.G., Clomburg, J.M., Conzalez, R. (2018). 'Synthetic pathway for the production of olivetolic acid in *Escherichia coli'*, *ACS Synthetic Biolpgy*, 7, article 8. Doi: 10.1021/acssynbio.8b00075

Taura, F., Morimoto, S., Shoyama, Y., Mechoulam, R. (1995). 'First direct evidence for the mechanism of delta.1-tetrahydrocannabinolic acid biosynthesis', *Journal of American Chemistry Society*, 117(38), pp. 9766-9767. Available at: https://pubs.acs.org/doi/pdf/10.1021/ja00143a024 (Accessed 29/08/2024)

Taura F, Morimoto S, Shoyama Y. (1996). 'Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa* L. biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid', *Journal of Biological Chem*istry, 271, pp 17411-17416. Doi: 10.1077/jbc.271.29.17411

Taura, F., Sirikantaramas, S., Shoyama, Y., Yoshikai, K., Shoyama, Y., &Morimoto,S. (2007). 'Cannabidiolic-acid synthase, the chemotype determining enzyme in the fiber-type *Cannabis sativa*', *FEBS Letters*, 581, pp. 2929-2934. Doi: 10.1016/j.febslet.2007.05.043

The *Arabidopsis* Genome Initiative. (2000). 'Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana'*, *Nature*, 408, pp. 796-815. Doi: 10.1038/35048692

Thichak, S., Natakankitkul, S., Chansakaow, S., Chotechuang, N. (2011)**.** 'Identification of drug-type and fiber-type hemp (*Cannabis sativa* L.) by multiplex PCR', *Chiang Mai Journal of Science,* 38(4), pp. 608-618. Available at: https://epg.science.cmu.ac.th/ejournal/ (Accessed 15/08/2025).

Tomas, B.F, ElSohly, M.A. (2015). 'The analytical chemistry of *Cannabis:* quality assessment, assurance and regulation of medicinal marijuana and cannabinoid preparations', *Elsevier Amsterdam1st edition*. Available at: https://www.amazon.com/Analytical-Chemistry-Cannabis-Cannabinoid-Preparations/dp/0128046465 (Accessed 13/03/2024)

Todaro, B. (2012). 'Cannabinoids in the treatment of chemotherapy induced nausea and vomiting', *Journal of the National Comprehensive Cancer Network*, 10, pp. 487-492. Doi: 10.6004/jnccn.2012.0048

Tucci, S.A., Halford, J.C.G., Harrold, J.A., Kirkham, T.C. (2006). 'Therapeutic potential of targeting the endocannabinoids: Implications for the treatment of obesity, metabolic syndrome, drug abuse and smoking cessation', *Current Medicinal Chemistry*, 13(12), pp. 2669-2680. Doi: https://doi.org/10.2174/092986706778201512

Turner, C.E. & ElSohly, M.A. (1981). 'Biological activity of cannabichromene, its homologs and isomers', *Journal of Clinical Pharmacology*, 21, pp. 283-291. Doi: 10.1002/j.1552-4604.1981.tb02606.x

Van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., Page, J.E. (2011). 'The draft genome and transcriptome of *Cannabis sativa*', *Genome Biology* 12: R102. Doi: 10.1186/gb-2011-12-10-r102

Viehmannová, I., Cusimamani, E.F., Bechyne, M., Vyvadilová, M., Greplová, M. (2009). '*In vitro* induction of polyploidy in yacon (*Smallanthus sonchifolius*)', *Plant Cell,Tissue and Organ Culture*, 97, pp. 21-25. Doi: 10.1007/s11240-008-9494-6.

Wilkinson, J.D. & Williamson, E.M. (2007). 'Cannabinoids inhibit human keratinocyte proliferation through a non-CB1/CB2 mechanism and have a potential therapeutic value in the treatment of psoriasis', *Journal of Dermatological Science*, 45(2), pp. 87-92. Doi: 10.1016/j.jdermsci.2006.10.009

Xiao, T. & Zhou, W. (2020). 'The third-generation sequencing: the advanced approach to genetic diseases', *Translational Pediatrics*, 9(2), pp. 163-173. Doi: 10.21037/tp.2020.03.06.

Xie, Z., Mi, Y., Kong, L., Gao, M., Chen, S., Chen, W., Meng, X., Wei Sun, W., Shilin Chen, S., Xu, Z. (2023). '*Cannabis sativa*: origin and history, glandular trichome development, and cannabinoid biosynthesis', *Horticulture Research,* 10. Doi: 10.1093/hr/uhad150

Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D., Ju, Y., Yao, P., Li, X., Xie, K., Zhang, J., Wang, J., Liu, F., Ma, W., Shopan, J., Zheng, H., Mackenzie, S.A., Zhang, M. (2016). 'The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection', *Nature Genetics*, 48, pp. 1225-1232. Doi: 10.1038/ng.3657

Yuan, Y., Bayer, P.E., Batley, J., Edwards, D. (2017). 'Improvements in genomic technologies: Application to crop genomics', *Trends in Biotechnology*, 35 (6), pp. 547-558. Doi:10.1016/j.tibtech.2017.02.009

Yoshikai, K.; Taura, T.; Morimoto, S.; Shoyama, Y. (210). 'DNA encoding cannabidiolate synthase', Patent in Japan JP 2000-78979A, DDBJ/EMBL/GenBank Database Accession Numbers E55107

Yuan, Y., Bayer, P.E., Batley, J., Edwards, D. (2017). 'Improvements in genomic technologies: Application to crop genomics', *Trends in Biotechnology*, 35, pp. 547-558. Doi: 10.1016/j.tibtech.2017.02.009

Zerbino, D. R., & Birney, E. (2008). 'Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs', *Genome Research, 18*(5), pp. 821-829. Doi: 10.1101/gr.074492.107.

Zuardi, A.W. (2006). 'History of *Cannabis* as a medicine: A review', *Brazilian Journal of Psychiatry*, 28(2), pp. 153-157. Doi: 10.1590/S1516-44462006000200015

Chapter Two: Sequencing, assembly and annotation of the *Cannabis sativa* Swazi Gold genome

2.1. Abstract

Cannabis sativa produces 125 distinct cannabinoids, the most abundant being delta-9-tetrahydrocannabinol (Δ^9- THC), cannabidiol (CBD), cannabichromene (CBC) and cannabigerol (CBG). These cannabinoids have demonstrated therapeutic potential in the treatment of various medical and psychological conditions; however, cannabinoid biosynthesis is not well understood. Here we report the draft annotated genome of the African landrace *C. sativa* Swazi Gold (SG). The genome was sequenced using both long-read (Pacific Bioscience) and short-read (Illumina) sequencing data. The genome was assembled using a range of short-read and hybrid assemblers. The performance of both short-read and hybrid assemblers was evaluated to obtain an optimal genome assembly. The optimal *de novo* genome assembly was produced using MEGAHIT with both short and long reads. Subsequent reference-guided assembly was undertaken using Rebaler and RaGOO. The final assembled genome comprised of 18 scaffolds, with a total size of 764.62 Megabases and a G+C content of 34.04% and a BUSCO completeness score of 97.0%. Structural and functional annotation of the assembled genome identified a total of 45,142 protein coding genes, 648 tRNAs and 907 rRNAs, while repetitive elements encompassed a predicted 77.19% of the genome. The draft genome of *C. sativa* SG represents the first genome of an African landrace, allowing for a better understanding of the genetics underlying cannabinoid biosynthesis in an African landrace, which could in turn be applied for biotechnological exploitation and comparative genomic analyses.

2.2. Introduction

Plant genome sequencing began with the publication of the first complete nuclear genome of *Arabidopsis thaliana* (*Arabidopsis* Genome Initiative, 2000). Since then, plant genome research has escalated rapidly, primarily driven through major projects including the Earth Biogenome project (EBG; https://www.earthbiogenome.org/); (Lewin *et al.,* 2022) and the 1000 Plant genome (1K) project (Leebens-Mack *et al.,* 2019). The genomes of >3,000 plant species from >1,000 families have been sequenced through these projects (Lewin *et al.,* 2022; Schwacke *at al*., 2025). However, this still only represents 0.1-0.2% of known plant species (Marks *et al.,* 2021).

A plant of significant research interest is *Cannabis sativa*, which produces a wide array of cannabinoids with therapeutic potential for treating inflammatory, neurological and

psychological disorders (Abrams *et al.,* 2007; ElSohly *et al.,* 2017; Liang *et al.,* 2004). A well-assembled genome could provide key information to further our understanding of the cannabinoid biosynthetic pathways and uncover the multifunctional applications of the plant (Gao *et al.,* 2020). Currently, twenty-five *C. sativa* genome assemblies are publicly available on the National Centre for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/). However, only a few of the available genomes are of chromosome-level quality, with most assemblies containing significant gaps. This limits their utility for downstream genetic and metabolic studies.

Poor *C. sativa* genome assemblies can primarily be attributed to a combination of the high genome repeat content associated with this species, as well as limitations imposed by the various technologies used for genome sequencing (Pisupati *et al.,* 2018). Additional notable challenges of plant genome assembly include the polyploid nature of plants and the large genome size (Neale *et al.,* 2014; Nystedt *et al.,* 2013, Salman-Minkov *et al.,* 2016).

The introduction of next generation sequencing (NGS) technologies provided a major step forward in the field of plant genome assembly (Michael & VanBuren, 2015). NGS platforms including 454, Illumina, SOLiD and Ion Torrent have all found application in the sequencing and construction of plant genomes (Bentley *et al.,* 2008; Changsheng *et al.,* 2018; Margulies *et al.,* 2005; McKernan *et al.,* 2009). However, approaches that rely solely on these NGS technologies still have shortcomings due to their inherently short reads, limiting their ability to cover the full range of the repetitive regions in plant genomes, resulting in fragmented assemblies, collapsed and unordered contigs (Changsheng *et al.,* 2018).

Third generation sequencing (TGS) technologies were developed to bridge many of these shortcomings, including Pacific Biosciences (PacBio) sequencing (Changsheng *et al.,* 2018; Xiao & Zhou, 2020). PacBio allows for the production of complete complex plant genome assemblies at the chromosome level (Changsheng *et al.,* 2018). However, the cost of long-read methods is still considerably high and continue to have notable pitfalls including lower throughput, the demand for very high-quality DNA and higher error rates (~10%) (Li & Schatz, 2025; Zhao *et al.,* 2023).

Hybrid genome assembly is often used to produce good quality plant genome assemblies (Bombarely *et al.,* 2016; Yang *et al.,* 2016). This approach utilises short-reads for initial assembly, followed by the use of long-reads for gap closure and scaffolding (Bombarely *et al.,* 2016; Yang *et al.,* 2016). Multiple hybrid genome assembly tools such as MaSuRCA,

HybridSPades and DBG2OLC, have been developed that can co-assemble short and long-read data (Chen *et al.,* 2020; Ye *et al.,* 2016; Zimin *et al.,* 2013). This hybrid approach allows for the production of plant genome assemblies with higher contiguity and accurate structural resolution (Cuperus and Questech, 2020).

In this study, we have produced a draft genome for the African landrace *C. sativa* Swazi Gold (SG) using both long- and short-read technologies. This landrace is classified as a drug-type chemotype and is primarily cultivated for its potent psychoactive effects and Δ^9-THC content (16-18%) (Hartsel *et al.,* 2016). The draft genome was achieved using an array of short-read and hybrid assemblers paired with reference-guided assembly. The draft genome for SG was subsequently structurally and functionally annotated for genomic analysis. The availability of this genome aids in improving our understanding of *Cannabis* chemotypes, cannabinoid biosynthesis and the diverse applications of this plant.

2.3. Methods and Materials

2.3.1. Plant material

All *C. sativa* SG plant material, in the form of fresh leaves and trichomes, was provided by The Root (Pty) Ltd, Pretoria, South Africa.

2.3.2. DNA Extraction

2.3.2.1. DNA extraction for Illumina sequencing

Total genomic DNA (gDNA) was extracted from *C. sativa* SG using the Zymo Quick-DNA Plant/Seed Kit (Zymo Research, USA) according to the protocol of the manufacturer. This kit produces clean, inhibitor free, high-quality, double-stranded gDNA that is suitable for Illumina sequencing (Illumina, 2024). In brief, freeze dried plant material was ground at room temperature using a sterile pestle and mortar and 150 mg was added to a ZR BashingBead™ lysis tube. BashingBead™ Buffer (750 µl) was added, and the sample was processed at maximum speed for 40 min using a Disruptor Genie (Scientific Industries Inc, USA) bead beater. The sample underwent centrifugation at 10,000 x *g* for 2 min, after which the supernatant was transferred to a Zymo-Spin™ III-F Filter in a collection tube. This was followed by a second round of centrifugation (8,000 x *g* for 1 min). The filter was discarded

and 1,200 µl Genomic Lysis Buffer was added to the filtrate and mixed. The mixture (800 µl) was transferred to a Zymo-Spin™ IICR Column in a collection tube and centrifuged (10,000 x *g* for 1 min). The flowthrough was discarded and the centrifugation step repeated. Next, 200 µl of Pre-Wash Buffer was added to the column and centrifuged (10,000 x *g* for 1 min). A total of 500 µl gDNA Wash Buffer was then added, followed by centrifugation (10,000 x *g* for 1 min). Finally, the column was transferred to a clean microcentrifuge tube and 100 µl DNA Elution Buffer was added and centrifuged (10,000 x *g* for 30 s) to elute the gDNA. The gDNA was stored at −20 °C prior to sequencing.

The integrity of the extracted gDNA was assessed using a 0.6% (w/v) agarose gel supplemented with SafeView™ Classic (Applied Biological Materials Inc., Canada). The concentration of the extracted gDNA was determined using the Nanodrop® ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA). The gDNA was submitted to Molecular Research LP (MR DNA, USA) for Illumina short-read sequencing.

2.3.2.2. DNA extraction for PacBio sequencing

Total gDNA for PacBio sequencing was extracted using a phenol/chloroform extraction protocol (Cullings, 1992; Doyle & Doyle, 1987). This protocol was selected over conventional bead bashing kits as PacBio sequencing requires high-quality, high-molecular-weight gDNA with minimal shearing to generate the characteristic long reads (Pacific Biosciences of California Inc. 2021). Freeze dried plant material was ground at room temperature using a sterile pestle and mortar, and ~ 100 mg was suspended in 500 µl of CTAB buffer (2% (w/v) cetyl trimethylammonium bromide, 1% (v/v) polyvinylpyrrolidone, 100 mM Tris-HCl, 1.4 M NaCl, 20 mM EDTA), and incubated at 60 °C for 30 min. Samples then underwent centrifugation at 14,000 x *g* for 10 min, after which the supernatant was removed and treated with 5 µl RNase A solution (10 mg/mL) (Thermo Fisher Scientific, USA) at 37 °C for 20 min to remove unwanted RNA (Doyle & Doyle, 1987). Nucleic acids were organically extracted using phenol-chloroform-isoamyl alcohol (25:24:1), which also promotes the removal of proteins and other contaminants from the nucleic acid solution through phase separation (Sambrook & Russell, 2001). Centrifugation (14,000 x *g* for 2 min) was performed to separate the phases with the DNA in the aqueous phase and cellular debris and lipids in the organic phase (Cullings, 1992; Doyle & Doyle, 1987). The DNA was precipitated using 0.7 volume 100% isopropanol and incubated at -20 °C for 15 min. The precipitated DNA was pelleted via

centrifugation (14,000 x *g* for 10 min). The DNA pellet was washed with 70% (v/v) ice-cold ethanol to remove any excess salts, followed by the evaporation of residual ethanol using a heating block. The DNA was resuspended in 20 µl of TE buffer (10 mM Tris, 1 mM EDTA, pH 8) and stored at −20 °C for sequencing.

The integrity and concentration of the extracted gDNA were assessed as described previously (section 2.3.1.1). The gDNA was submitted to Inqaba Biotech (Pty) Ltd (South Africa) for PacBio sequencing.

2.3.3. Library preparation and sequencing

2.3.3.1. Illumina library preparation and sequencing

Libraries were prepared using the Nextera DNA Library Prep Kit (Illumina, USA). Briefly, the manufacturers library preparation protocol entailed a five step workflow: 1) ~50 ng of the high-quality gDNA was enzymatically fragmented and simultaneously tagged with partial Nextera adapter sequences through Tn5-mediated tagmentation (Illumina Adapter Sequences, 2021); 2) full-length adapters containing dual indices and flow cell binding sites were added using limited-cycle PCR; 3) the gDNA library was purified using AMPure XP beads; 4) the gDNA library underwent quality analysis using Bioanalyzer and Qubit fluorometer and 5) the gDNA library was spiked with 1–2% PhiX Control v3, which served as a positive control for sequencing (Illumina, 2021). Sequencing of the prepared gDNA library was then performed using the Illumina NovaSeq 6000 platform with the 2 × 250 bp paired-end (PE)sequencing configuration for an expected yield of 70 million reads

2.3.3.2. PacBio library preparation and sequencing

The SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, USA) was used for library preparation. The workflow for library preparation involved eight steps: 1) the quality of the gDNA was determined; 2) the gDNA was fragmented using a Covaris g-TUBE; 3) the concentration of the fragmented gDNA was adjusted using AMPure magnetic beads; 4) gDNA repair was performed to fix damaged nucleotide bases and remove single-strand overhangs (to generate clean fragmented gDNA), followed by A-tailing to prepare for adapter ligation; 5) blunt-end hairpin SMRTbell adapters were ligated to both ends of the double-stranded gDNA fragments; 6) the adapter-tagged gDNA was purified using exonuclease; 7) primers

were annealed to the gDNA and 8) the resultant gDNA libraries were subjected to a final purification step to ensure they were contaminant free and ready for submission to the sequencer (Kong *et al.,* 2017; Pacific Biosciences of California Inc. 2021).

2.3.4. Genomic sequence quality control and trimming

The quality of the raw sequence data (Illumina short-read and PacBio long-read) was assessed using FastQC v. 0.11.9 (Andrews, 2010) and MultiQC v. 1.12 (Ewels *et al.,* 2016). Based on the FastQC and MultiQC outputs, the sequence reads were trimmed of low-quality bases and adapter sequences using TrimGalore v. 0.6.7 (Krueger, 2015) with Cutadapt v. 3.5 (Martin, 2011). Low quality bases were removed using two different Phred (Q-score) values: 20 and 25. Phred score provide a value of confidence that a DNA base has been correctly identified, with a higher value indicating a decreased chance of error (Shrestha *et al.,* 2014). A cutoff of Q20 reflects an error rate of 1 in 100 (1%) bases, while a cutoff of Q25 has an error rate of 1 in 316 (0.316%) bases (MacManes, 2014). Subsequent *de novo* assembly with MEGAHIT v. 1.2.9 (Li *et al.,* 2015) showed that better assembly was achieved when trimming with a Q-score of 20. Consequently, only the Q20 trimmed dataset was used for subsequent genome assembly (Bradnam *et al.,* 2013).

2.3.4. Short-read *de novo* genome assembly

The trimmed Illumina short-read dataset was assembled using three different algorithms, namely Assembly by Short Sequences (AbySS) v. 4.4.1 (Van Bakel *et al.,* 2011), MEGAHIT v. 1.2.9 (Li *et al.,* 2015) and String Graph Assembler (SGA) v. 0.10.15 (Simpson & Durbin, 2012).

AbySS is a De Bruijn graph-based (DBG) assembler that uses a distributed-memory parallel system, making it ideal for large and complex short-read datasets (Van Bakel *et al.,* 2011). This parallel memory maintains computer efficiency; however, it often requires follow up refinement and polishing tools for repeat-rich sections and complex genomes (Jackman *et al.,* 2017). AbySS commences by performing K-mer counting and graph construction, followed by graph simplification with the removal of low-abundance K-mers, tips, bubbles and erroneous branches (Van Bakel *et al.,* 2011). The simplified graph is then traversed to construct contigs, followed by gap filling and scaffolding to join contigs into longer scaffolds (Simpson *et al.,*

2009). For this study, AbySS was evaluated with various K-mer lengths (range 31-191) and two distinct K-mer coverage (KC = 2-3) values.

MEGAHIT is a fast and memory-efficient succinct De Bruijn Graph (sDBG)-based assembler which was developed for assembling large and complex metagenomic or single-genome datasets, to produce high coverage and high-quality genomes (Li *et al.,* 2015). The sDBG algorithm allows it to efficiently manage large volumes of sequencing data, while having reduced computational requirements to produce high quality assemblies. This makes it well suited for the assembly of genomes with highly repetitive regions such as plant genomes (Li *et al.,* 2016). The assembly pipeline incorporates four steps: 1) an error correction and K-mer filtering step where high quality K-mers are selected for graph construction; 2) construction of an sdBG using the selected K-mers; 3) generation of an integrative graph assembly using multiple K-mer sizes to fill complex regions and extend contigs; 4) resolution of gaps and bubbles in the graph and contig generation (Chikhi & Rizk, 2013; Li *et al.,* 2015; Peng *et al.,* 2012; Zerbino & Birney, 2008). For this study, MEGAHIT v. 1.2.9 was evaluated with K-mer lengths using default (21,29,39,59,79,99,119,141) parameters, followed by the evaluation of narrower K-mer ranges.

SGA is an overlap-based string graph model assembler, differing from most *de novo* assemblers, which utilise a DBG-based algorithm (Simpson & Durbin, 2012). It was designed to require low memory and have a reduced runtime (Schatz *et al.,* 2010). This reduced runtime is achieved through parallelization where overlaps between sequence reads are computed simultaneously across central processing unit (CPU) cores. Briefly, its workflow comprises five steps: 1) initial preprocessing of the raw sequence data to trim low-quality bases and filter short or low-complexity reads; 2) the construction of an Ferragina–Manzini index (FM-index) with the read data for fast, memory efficient overlap detection; 3) building of a string graph of the read overlaps to capture the genome structure; 4) simplification of the resultant string graph to remove bubbles and errors and 5) traversal of the simplified string graph to generate contigs (Schatz *et al.,* 2010). For this study, SGA was run for *de novo* assembly using its default parameters.

2.3.5. Evaluation of *de novo* genome assemblies

All *de novo* short-read genome assemblies were assessed using QUAST-LG v. 5.2.0 (Gurevich *et al.,* 2013), to determine which genome assembler gave the optimal assembly. The parameters

used to define this optimal assembly included the N50 value, the number and length of contigs, the length of the largest contig, overall genome size and the genome fraction.

The contig N50 value is a measure of the extent of fragmentation of an assembled genome (Bradnam *et al.,* 2013). This value is calculated by ordering the fragment lengths by size and, starting with the largest, adding the next smallest fragment until the sum of the fragments is greater than 50% of the total length of the assembly (Juahal & Newcomb, 2021). A genome with fewer contigs indicates a more contiguous and complete assembly. This suggests that the genome assembly is of a high quality as more sequences have been combined into longer continuous pieces by the assembler. In contrast, a high number of contigs suggests a more fragmented assembly which could be a result of sequencing errors, insufficient coverage or highly repetitive nucleotide sequences (Yandell & Ence, 2012).

Based on available *C. sativa* genomes, an expected size of ~ 700-900 Mb is predicted for this species. Thus, we evaluated the sizes of the different SG assemblies this expected *C. sativa* genome size.

Lastly, the genome fraction value indicates the percentage of the nucleotide bases in the draft assembly that align with the reference genome (Gurevich *et al.,* 2013). A higher genome fraction indicates a more complete and more comprehensive genome assembly, with a genome fraction value over 90% indicating a high-quality genome (Abbas *et al.,* 2024). The draft SG assemblies were aligned against the reference genome of *C. sativa* (GCF_029168945.1, Cultivar: Pink Pepper) using Minimap2 (Li, 2018) and ambiguous and redundant alignments were filtered out by QUAST-LG (Gurevich *et al.,* 2013).

2.3.6. Hybrid assembly (Illumina short-read and PacBio long-read)

The PacBio long-read data and Illumina short-read datasets were assembled using three hybrid genome assemblers, namely De Bruijn Graph to Overlap-Layout-Consensus (DBG2OLC) v. 1 (Ye *et al.,* 2016), MaSurCa v. 4.1.0 (Zimin *et al.,* 2013) and MEGAHIT v. 1.2.9. (Li *et al.,* 2015).

The DBG2OLC hybrid assembly tool produces an assembly by combining the PacBio long-reads for a backbone and subsequently using the Illumina short-reads for error correction, an approach that is particularly useful for highly repetitive genomes (Nystedt *et al.,* 2013). The major advantages of DBG2OLC include its compact representation of the long-reads (which

allows for efficient alignments); its ability to skip errors at a base level and detect structural errors which is followed by subsequent polishing of structurally correct reads (Ye *et al.,* 2016). This tool produces hybrid assemblies in five major steps: 1) DBG is built using the short read datasets, and contigs are constructed (Ye *et al.,* 2016; Zerbino & Birney, 2008); 2) the contigs are then mapped to long reads so that each long read is represented by a sequence of contig identifiers (Ye *et al.,* 2016); 3) the compressed long reads are then cleaned and reads with structural errors are removed using multiple sequence alignment (Ye *et al.,* 2016; Bertrant *et al.,* 2019); 4) an overlap graph is then constructed with the cleaned compressed reads; 5) finally, the long reads are uncompressed, chained together and converted to contigs (Ye *et al.,* 2016). In this study, DBG2OLC was run using the default parameter settings, which included a K-mer length of 17 for DBG construction, K-mer coverage threshold of 2, minimum overlap of 20 and an adaptive threshold of 0.005.

MaSurCa is a software tool designed to produce high quality assemblies or improve assemblies through gap closing, specifically for large and complex genomes using NGS data (Zimin *et al.,* 2013). It is capable of utilising both Illumina short-read and PacBio long-read data to produce a hybrid *de novo* assembly. Uniquely, this tool has a built-in error-correction feature allowing the direct use of raw untrimmed data (Zimin *et al.,* 2013). Its workflow includes five steps: 1) initial preprocessing, which includes pre-assembly correction of raw reads (short- and long-reads); 2) merging of corrected short-read data with long-read data to produce "super-reads" with high-quality Phred scores; 3) combining of "super-reads" to construct contigs using the overlap-layout-consensus approach; 4) scaffolding of the contigs into longer sequences using paired-end read information and 5) post-assembly polishing to fix errors and fill gaps (Zimin *et al.,* 2013). MaSurCa was run using the default parameter settings to produce a hybrid *de novo* assembly.

While MEGAHIT v. 1.2.9 is designed for assembly of short-read genomic data, long-reads can be added as single-end reads (option *-r*) alongside the paired-end short-read data (Li *et al.,* 2015). As such, assembly was first undertaken using the default K-mer lengths (21,29,39,59,79,99,119,141), followed by individual K-mer lengths (21, 41, 61, 81, 101, 121, 141, 161, 181, 201, 221) covering the full range of allowable K-mers lengths for MEGAHIT. An optimal K-mer range of 121-161 was identified and narrower K-mer length analysis was conducted (131, 151, 155 and 165). The final assembly was generated with a K-mer range 141-181 with K-mer steps of 4.

The quality (N50 value, number of contigs, genome assembly size and fraction) of the hybrid SG assemblies were determined using QUAST-LG v. 5.2.0 (Gurevich *et al.,* 2013) as outlined previously (section 2.3.5.), and the optimal *de novo* assembly was selected for subsequent reference-guided assembly and refinement approaches.

2.3.7. Reference-guided assembly

Reference-guided assembly was conducted due to the large size of the SG MEGAHIT hybrid genome assembly (3.77Gb) and large number of contigs produced. Two reference-guided assembly pipelines namely, RaGOO v. 1.1 (Alonge *et al.,* 2019) and Rebaler v. 0.2.0 (Wick *et al.,* 2019), were evaluated to improve genome contiguity. In both cases, the genome of *C. sativa* (GCF_029168945.1, Cultivar: Pink Pepper) was used as the reference.

RaGOO aligns draft genome contigs to the reference genome to improve the accuracy and contiguity of an assembly on the chromosome level (Alonge *et al.,* 2019). Initially the draft contigs are aligned to the reference genome using Minimap2 (Li, 2018), followed by the clustering of contigs, before ordering and orienting of the contigs linearly to produce the best alignment. The remaining chimeric contigs (misassembled contigs) are split at breakpoint locations and finally, an improved, structurally accurate reference-guided scaffolded assembly is produced (Alonge *et al.,* 2019). Unaligned draft assembly contigs are retained. To ensure that unaligned draft assembly contigs were not SG-specific elements, seqtk v. 1.3-r106 (Li, 2018) was used to filter out contigs < 1000 bp in size. The remaining contigs were annotated using Modular Open-Source Genome Annotator (MOSGA) v. 2.1.5 (Martin *et al.,* 2020).

Rebaler v. 0.2.0 (Wick *et al.,* 2019) allows for long-read genome reconstruction by first aligning the assembly to the reference genome using Minimap2. This is followed by refinement of the assembled draft genome with multiple rounds of Racon (Vaser *et al.,* 2017) for polishing. A key issue with Rebaler is that it inserts reference sequence in assembly gaps in the draft genome (Wick *et al.,* 2019). To counteract this, the Illumina short-read and PacBio long-read data were mapped against the Rebaler-guided *C. sativa* SG assembly using the BWA-MEM algorithm (part of the alignment via Burrows-Wheeler transformation (BWA) v. 0.7.17-r1188 package) (Li & Durbin, 2009) and Minimap2 (Li, 2018), respectively. The resultant .sam files were combined, converted to .bam format, sorted and indexed using samtools v. 1.13 (Danecek *et al.,* 2021). The coverage depth was determined from the .bam alignment file using the *depth* command of samtools v. 1.13. Zero-coverage regions were masked (replaced by Ns) in the

Rebaler-guided *C. sativa* SG draft assembly using bedtools v. 2.30.0 (Quinlan & Hall, 2010) with the *maskfasta* command. As polishing of the genome was done using Racon (Vaser *et al.,* 2017) the remaining gaps corresponded to regions with no read coverage and thus, no further gap closing was possible. Therefore, the Rebaler-guided assembly was directly subjected to annotation.

2.3.8. Refinement of the *C. sativa* SG RaGOO-guided genome assembly

2.3.8.1. Long-read gap closing

Gaps in the RaGOO-guided SG genome assembly were closed using several long-read gap closing tools. Long-read gap closure was undertaken sequentially using LR_Gapcloser (Xu *et al.,* 2018), MaSurCa v. 4.1.0 (Zimin *et al.,* 2013) and TGS-GapCloser v. 1.2.1 (Xu *et al.,* 2020). LR_Gapcloser converts the long-read data into tags of 300 bp and aligns them to the draft assembly using BWA-MEM, followed by tag classification, gap bridging and iterative refinement (Xu *et al.,*2020). MaSurCa uses the SAMBA (Zimin & Salzberg, 2022) tool to accurately improve genome contiguity through scaffolding and gap filling. SAMBA can fill in gaps within the genome sequence scaffolds using additional long-read data (Zimin & Salzberg, 2022). TGS-GapCloser v. 1.2.1 was developed to use low-coverage raw long-read data to improve genome completeness and contiguity with accuracy (Xu *et al.,*2020). It was used to align the long reads to the RaGOO-guided assembly using Minimap2, followed by polishing and gap closing.

2.3.8.2. Illumina short-read gap closing

The resulting long-read gap closed RaGOO-guided assembly was further subjected to short-read gap closing with the short-read Illumina data. Short-read gap closing was conducted sequentially using short-read gap closers AbySS_sealer (Paulino *et al.,* 2015) and GapCloser v. 1.12 (Luo *et al.,* 2012).

AbySS_sealer uses a bloom filter SdBG approach and works efficiently on very large genomes (Jackman *et al.,* 2017). The bloom filter improves memory efficiency by storing parts of the constructed K-mers, checking if each K-mer exists whilst efficiently finding the paths within the DBG to fill gaps. GapCloser v. 1.12 (Luo *et al.,* 2012) identifies gaps as series of Ns in the

assembly and maps the short reads using a K-mer-based approach, and bridges gaps using abundant pair relationships (Luo *et al.,* 2012).

The efficacy of both long- and short-read gap closing was evaluated using QUAST-LG v. 5.2.0 as outlined previously (section 2.3.5.).

2.3.8.3. Polishing

To correct any single-base substitution, deletion and insertion errors in the genome (Vaser *et al.,* 2017; Walker *et al.,* 2014), the gap-closed RaGOO-guided assembly (long- and short-read gap closed) underwent a round of polishing using Pilon v. 1.23 (Walker *et al.,* 2014). The draft assembly was split into 18 individual scaffolds, and the Illumina short-read data was aligned against the scaffolds using BWA-MEM (José *et al.,* 2015). The aligned data was used to identify and correct errors present in the draft assembly, generating more accurate and continuous sequences (Vaser *et al.,* 2017; Walker *et al.,* 2014). The final polished RaGOO-guided assembly and Rebaler-guided assemblies (section 2.3.7.) were subjected to QUAST-LG v. 5.2.0 to assess assembly quality (section 2.3.3.1).

2.3.9. Genome annotation

The Rebaler and RaGOO reference-guided *C. sativa* SG genome assemblies (section 2.3.4.) were evaluated for completeness using BUSCO v. 3.0 (Waterhouse *et al.,* 2018) with the default settings. BUSCO is a bioinformatics tool that assesses genome assembly completeness using conserved, single-copy orthologous genes (SCOs) (Simão *et al.,* 2015). BUSCO uses lineage-specific datasets for comparisons between a wide range of species, these datasets are comprised of orthologous genes that are highly conserved across the different evolutionary lineages (Waterhouse *et al.,* 2018). These datasets allow for the identification of complete, fragmented, duplicated and missing genes in the assembled genome and like-for-like comparisons of genomes. The assemblies underwent comparison against the eudicots_odb10 lineage dataset (n = 2326 species) using BUSCO v. 3.0 (Waterhouse *et al.,* 2018) to assess genome completeness. This eudicots_odb10 lineage dataset is specifically tailored for flowering plant genomes (Waterhouse *et al.,* 2018). A BUSCO completeness score of above 90% is the general benchmark used in literature for a high-quality reference-level assembly, including the EBP (EBG; https://www.earthbiogenome.org/) and Darwin Tree of Life Project

(Darwin Tree of Life Project Consortium, 2022; https://www.darwintreeoflife.org/) (Lewin *et al.,* 2022). BUSCO completeness scores ranging from 70-90% have been recorded for draft genome level assemblies (BMC Genomics, 2024; Michael & VanBuren, 2020).

The Rebaler and RaGOO reference-guided *C. sativa* SG genome assemblies (section 2.3.4.) were structurally and functionally annotated using MOSGA v. 2.1.5 pipeline (Martin *et al.,* 2020). MOSGA utilises various prediction tools, producing results in formats suitable for NCBI submission and genome browser analysis (Martin *et al.,* 2020). First, repeat content was determined using REpeat Detector (RED), which allows for *de novo*, reference-free repeat identification (Girgis, 2015). RED results were used to soft-mask the genome sequence for subsequent annotation. The advantages of this tool include the detection of unknown repeats without relying on available pre-existing libraries; however, the output is limited to generic repeat content without classification into families (e.g. DNA transposons, LTRs) (Girgis, 2015). As such, a repeat library was generated using the reference genome of *C. sativa* (GCF_029168945.1) using RepeatModeler v. 2.0.7 (Flynn *et al.,* 2020). RepeatModeler utilises multiple tools for the *de novo* identification of repeat families namely, RECON (Bao & Eddy, 2002) and RepeatScout (Price *et al.,* 2005). Additionally, RepeatModeler uses Tandem Repeats Finder (TRF) (Benson, 1999) for detecting tandem repeats within the assembly and RepeatClassifier (Flynn *et al.,* 2020) for repeat library sequence annotation and refinement. The resultant library was used to predict and identify repetitive elements in the *C. sativa* SG assemblies using RepeatMasker v. 4.2.0 (Smit *et al.,* 2013).

Transfer RNA elements (tRNAs) were predicted using tRNAscan-SE2 v. 2.0 (Lowe & Chan, 2016), which uses a covariant model for tRNA gene and function prediction. This model allows for sensitive tRNA searches with a broad diversity of tRNA genes and detects mitochondrial tRNAs (Lowe & Chan, 2016). Ribosomal RNA elements (rRNAs) were predicted using BAsic rapid Ribosomal RNA Prediction (Barrnap) v. 0.9 (Seemann, 2018). Barrnap predicts rRNA genes by scanning the sequences using Hidden Markov Models (HMMs) which have been trained with known rRNA sequences. Through this process, conserved motifs and structural features are identified resulting in rRNA prediction files for downstream annotation analyses (Seemann, 2018).

Protein prediction was performed in MOSGA using AUGUSTUS (Stanke *et al*, 2004). AUGUSTUS allows for *ab initio* gene prediction, and it identifies protein coding genes as well as their exon-intron structure in assembly sequences (Brůna *et al.,* 2023). The gene models for *Vitis vinifera* (NCBI RefSeq assembly GCF_030704535.1, Cultivar: Pinot Noir 40024) served

as the gene model to predict genes, as this is the least distant reference genome with available gene models as predicted by MOSGA.

EggNOG-mapper (Huerta-Cepas *et al.,* 2017) and SWISS-PROT (Bairoch & Apweiler, 2000) were used for functional annotation and classification of the proteins predicted by AUGUSTUS. SWISS-PROT is a protein sequence database that provides annotation at a high level with minimal redundancy (Bairoch & Apweiler, 2000). Uniquely, this curated database is integrated through partnerships with additional databases including the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) (Cook *et al.,* 2017) and the Swiss Institute of Bioinformatics (SIB) (D'Auria, 2017), providing an integrated annotation approach. The eggNOG-mapper tool can annotate large proteins datasets by performing fast orthologue assignments against the eggNOG v. 5.0 database (Huerta-Cepas *et al.,* 2017).

2.4. Results and Discussion

2.4.4. DNA extraction

Total genomic DNA (gDNA) was extracted from *C. sativa* SG using the Zymo Quick-DNA Plant/Seed Kit. The gDNA had a concentration of 80.08 ng/µl and an A260/280 ratio of 1.92. An A260/280 ratio of between 1.8-2.0 is considered acceptable for downstream applications such as Illumina sequencing (Illumina, 2024; Glasel, 1995).

The phenol-chloroform extracted gDNA had a concentration of 135.3 ng/µl and an A260/280 ratio of 1.52. Although the achieved gDNA concentration was acceptable for PacBio sequencing, the low A260/280 ratio, which is indicative of protein or phenol contamination (Glasel, 1995), fell outside the acceptable range (1.8-2.0). Nevertheless, the phenol-chloroform extracted gDNA was sent for PacBio long-read sequencing.

2.4.5. Genomic sequence quality control

The *C. sativa* SG genome was initially sequenced on the Illumina NovaSeq 6000 platform using the 2 x 250 bp paired-end (PE) sequencing configuration. The Illumina sequencing generated a total of 94,211,563 PE reads and total yield of ~41 Gigabases (Gb), giving an expected genome coverage of 53x (Table 2.1). To improve the completeness and quality of the assembled SG genome, additional long-read PacBio sequencing was conducted. However, with

a total yield of 1.4 Gb, PacBio sequencing yielded a very low genome coverage of 1.7x (Table 2.1).

Quality control analysis of the Illumina raw short-read sequencing data using FastQC and MultiQC indicated that the per nucleotide base quality for nucleotides located at the beginning and central portion of the reads (base positions 1-200) exhibited high quality scores, while the terminal segments (bases 190-240) displayed a decline in quality (Appendix Figure A1). A Phred score cut-off of Q20 and Q25 was identified for the trimming of the Illumina raw short-read sequence data to remove low-quality base sequences. Trimming with the Q20 cutoff resulted in the removal of 0.2% (forward) and 1.0% (reverse) PE reads (Table 2.2). The Q25 cutoff resulted in the removal of 0.7% (forward) and 1.0% (reverse) PE reads (Table 2.2). The trimmed Illumina short-read data was reassessed using FastQC and MultiQC, which confirmed that the low-quality bases at the end of the reads and the Nextera adapters had been successfully removed, ensuring that the Illumina short-read sequence data was ready for genome assembly (Appendix Figure A2).

Quality control analysis of the PacBio raw long-read sequence data using FastQC and MultiQC revealed the very low genome coverage of 1.7x (Table 2.1). Given the low coverage of the PacBio data, trimming was not conducted to retain the maximum amount of the sequence data for the hybrid assembly and gap filling of the Illumina assemblies.

Table 2.1: MultiQC v. 1.12 and FAstQC v. 0.11.9 Illumina and PacBio raw sequence data analysis.

Sequence data	**Illumina paired-end (PE) reads**	**PacBio reads**
Number of reads	94 211 563	339 103
Average read length (bp)	215.9	3 981
Sequencing yield (bp)	40 674 974 38	1 350 203 405

Table 2.2: Illumina short-read sequence data results after trimming at Phred score cutoffs of Q20 and Q25 using TrimGalore v. 0.6.7 with Cutadapt v. 3.5.

Data (bp)	Raw paired-end (PE) reads	Q20		Q25	
		Forward	Reverse	Forward	Reverse
Total sequences	94,211,563	20,343,089,487	20,331,885,451	20,343,089,487	20,331,885,451
Quality trimmed	36	33,834,839	206,200,221	139,705,803	44,631,577
Retained	N/A	20,214,786,124	20,039,537,527	20,113,463,749	20,195,209,601

2.4.6. *de novo* Genome assembly

2.4.6.1. Short-read *de novo* genome assembly

The Illumina short-read data was assembled using three *de novo* pipelines: AbySS v. 4.4.1, MEGAHIT v. 1.2.9 and SGA v. 0.10.15. The largest genome assembly was obtained using MEGAHIT (3.77 Gb), ~ 4x the expected assembly size for *C. sativa* (700-900 Mb) (Table 2.3). In contrast, SGA produced an assembly (948.80 Mb) that more closely reflected the expected size for a *C. sativa* genome.

The MEGAHIT assembly comprised the largest number of contigs (Table 2.3), ~6x as many as obtained through assembly with SGA. The MEGAHIT assembly furthermore produced the largest contig, with the largest contig being 3.5x and 8x the size of the largest contig obtained with AbySS and SGA, respectively (Table 2.3).

The lowest N50 value was observed for the MEGAHIT assembly, while the highest N50 value was achieved with AbySS (Table 2.3).

The starkest difference between the assemblies was observed for the genome fraction (when compared to the reference genome of *C. sativa* (GCF_029168945.1, Cultivar: Pink Pepper). The highest genome fraction was observed for the MEGAHIT assembly (86.56%), while those of the SGA and AbySS assemblies were negligible (Table 2.3). This suggests that, while the SGA and AbySS assemblies more closely reflect the expected *C. sativa* genome in terms of

size, these assemblers are prone to miss-assembly, which would be expected to lower the achieved genome fraction (Gurevich *et al.,* 2013).

Overall, all assemblers produced an assembly comprising a substantial number of contigs (Table 2.3). A genome with fewer contigs indicates a more contiguous and complete assembly (Juahal & Newcomb, 2021), and as such, in order to improve genome contiguity and completeness, long-read sequencing data was generated from *C. sativa* SG and hybrid *de novo* genome assembly was performed.

Table 2.3: Genome metrics of assemblies produced on the Q20 trimmed Illumina short-read data with AbySS v. 4.4., MEGAHIT v. 1.2.9 and SGA v. 0.10.15.

Assembler	Genome size (Mb)	K-mer range	K-mer coverage	Genome fraction (%)	N50 (bp)	Contigs >1000 bp	Largest contig (bp)
AbySS	1,227.49	116	2	5.688	4,676	97,792	36,399
MEGAHIT	3,766.40	101–141 (step 2)	--	86.56	393	282,349	128,613
SGA	948.80	51	1	1.706	4,077	47,394	15,912

2.4.6.2. Hybrid *de novo* genome assembly

The PacBio long-read and Illumina short-read data were assembled using three hybrid *de novo* assemblers.

DBG2OLC and MaSurCa both yielded genome assemblies (4.6 and 105 Mb, respectively) that were far below the expected assembly for *C. sativa* (700-900 Mb) (Table 2.4). Furthermore, lower N50 values and genome fractions were observed for these assemblies compared to that produced by MEGAHIT. This was likely a result of DBG2OLC and MaSurCa relying on greater long-read sequencing depth than was achieved in this study. Long-read assembly for plant genome typically requires a coverage of 30-60x, with assembly of PacBio long-read data working well with a coverage of approximately 30x (Koren *et al*. 2017). The DBG2OLC assembler typically works optimally with ≥ 30x long-read coverage to generate high quality complex assemblies (Ye *et al*. 2016). Notably, hybrid assembly using the MEGAHIT assembler recommends a coverage of ~40-50x for plant genome projects (Li *et al.,* 2018).

However, the long-read PacBio sequencing data generated for this study gave a low expected genome coverage of 1.7x, resulting in little to no improvement to the Illumina data generated assemblies.

Table 2.4: Optimal hybrid *de novo* assemblies produced using DBG2OLC v. 1., MaSurCa v. 4.1.0, and hybrid MEGAHIT v. 1.2.9.

Assembler	Genome size (Mb)	K-mer range	Genome fraction (%)	N50 (bp)	Contigs >1000 bp	Largest contig bp
DBG2OLC	4.659	116	0.002	7,802	627	17,302
MaSurCa	105.603	--	4.241	4,456	42,088	106,400
Hybrid MEGAHIT	3766.397 (3.77 Gb)	101–141 (step 2)	11.909	4,759	282,349	128,613

Comparison of the hybrid assembly produced with MEGAHIT to the DGB2OLC and MaSurCa hybrid assemblies showed that MEGAHIT achieved the largest assembly, with the greatest number of contigs > 50,000 bp in size (six, compared to only one for MaSurCa) (Table 2.4), as well the largest contig produced (128,613 bp, compared to DBG2OLC – 17,302 bp and MaSurCa – 106,400 bp). Furthermore, the highest genome fraction was achieved with MEGAHIT (11.909%), more than double that achieved with MaSurCa (4.241%). The N50 value of the MEGAHIT assembly (4,759 bp) was higher than that achieved with MaSurCa (4,456bp), but lower than that produced with DBG2OLC (7,802 bp), albeit that the latter only comprised 627 contigs, while the MEGAHIT assembly comprises 9,261,818 contigs.

Comparison of the best hybrid assembly (MEGAHIT) to the assembly produced from the short-read data alone (MEGAHIT) showed that both assemblies had the same genome size (3.77 Gb), largest contig size and number of contigs greater than 1,000 bp. The short-read assembly (MEGAHIT) had the highest genome fraction (86.56%) however, the hybrid assembly (MEGAHIT) had a higher N50 value of 4,759 bp. The lower genome fraction achieved for the hybrid assembly (MEGAHIT) was likely due to miss-assembly which occurred when the long-read data was included.

As such, the hybrid assembly with MEGAHIT provided the optimal *de novo* assembly. However, this assembly was ~ 4-5x the expected size for a *C. sativa* genome, suggesting miss-assemblies within the SG draft genome and potential duplication errors introduced during assembly. The low N50 value (4,759 bp) and high number of contigs indicated that while the genome remained very fragmented and required gap filling and refinement. As such, to improve the contiguity and more accurately assemble the contigs into "chromosome" structures, reference-guided improvement of the MEGAHIT hybrid assembly was undertaken.

2.4.7. Reference-guided assembly

The MEGAHIT hybrid assembly was mapped against the reference genome of *C. sativa* (GCF_029168945.1, Cultivar: Pink Pepper) using the reference-guided assembly tools RaGOO v. 1.1 (Alonge *et al.,* 2019) and Rebaler v. 0.2.0 (Wick *et al.,* 2019).

The RaGOO and Rebaler assemblies both comprised of eighteen scaffolds reflecting the chromosomes in the reference assembly, with the Rebaler assembly being the largest (764.62 vs 436.89 Mb) (Table 2.5). The genome size of the Rebaler-based assembly was consistent with available *C. sativa* genomes on the NCBI database, (Appendix Table A1). Therefore, the Rebaler-based genome assembly was considered the better assembly, while the RaGOO assembly required further refinement.

A total of 233,400 unaligned contigs were retained from the draft assembly following RaGOO reference-guided assembly. These additional contigs (RAGOO additional) underwent gap closure and polishing with Pilon v. 1.23 as per the RaGOO assembly.

Table 2.5: Assemblies generated using two reference-guide assembly strategies RaGOO v. 1.1 and Rebaler v. 0.2.0 from the MEGAHIT hybrid genome assembly.

Reference-based assembler	Genome size (Mb)	Genome fraction (%)	N50 (bp)	Contigs >1000 bp	Largest contig bp
RaGOO	436.89	29.072	43,784,932	18	61,502,276
RaGOO additional	470.40	0.022	4,987	233,400	26,914
Rebaler	764.62	89.865	76,567,386	18	91,356,502

2.4.8. Refinement of the draft *C. sativa* SG genome assembly

The RaGOO-guided genome assembly underwent contig extension, scaffolding, gap closing, polishing and refinement. Long-read gap closing was performed using LR_Gapcloser (Xu *et al.,* 2018), MaSurCa v. 4.1.0 (Zimin *et al.,* 2013) and TGS-GapCloser v. 1.2.1 (Xu *et al.,* 2020). Only one tool, namely TGS-GapCloser v. 1.2.1 closed ~1.91% of the gap positions. The performance of the different tools likely reflects their ability to close gaps in low coverage sequence data, a known advantage associated with TGS-GapCloser (Xu *et al.,*2020).

ABySS-Sealer and GapCloser v. 1.12 were subsequently used sequentially for short-read gap closure. While ABySS-Sealer was unable to close any gaps, GapCloser successfully closed 78.16% of the gap positions, resulting in a final gap-closed assembly containing 1,098.43 Ns/ 100 kbp. The Ns/100 kb value is a measure of the level of fragmentation or gaps in the assembly which were introduced during scaffolding (Bradnam *et al.,* 2013; Rhie *et al.,* 2021).

Final genome polishing was performed using Pilon v. 1.23. This polishing reduced the overall size of the genome assembly by 4.46 Mb, resulting in a genome size of 438.36 Mb. Pilon also reduced the number of gaps in the assembly to a final value of 1057.78 Ns/100 kbp.

2.4.9. Comparison of the final SG genome assemblies

The Rebaler reference-guided assembly, refined RaGOO reference-guided assembly and the retained additional RaGOO assembly contigs were structurally and functionally annotated using the MOSGA pipeline (Smit *et al.,* 2013).

2.4.9.1. Genome metrics and genome completeness

The largest genome was obtained using the Rebaler reference-guided assembly tool (Table 2.6), which fell within the expected range for the genome size (700-900 Mb) for a complete *C. sativa* genome. By contrast, the refined RaGOO assembly and the additional RaGOO contigs, were each closer to half the size of the reference genome of *C. sativa* (GCF_029168945.1; Cultivar: Pink Pepper) (Table 2.6). All three annotations had similar G+C contents ranging from 34.04 to 35.69 (Table 2.6), which were consistent with the G+C content levels in available *C. sativa* genomes (Appendix Table A1). The Rebaler-guided assembly demonstrated a genome fraction of 89.87% and BUSCO completeness score of 97.0%. This BUSCO value meets the requirements of both an acceptable high-quality draft genome and a reference-grade genome (Bradnam *et al.,* 2013; Michael & VanBuren, 2020). In contrast, the additional RaGOO contigs achieved the lowest genome fraction (0.022%) and BUSCO completeness value (37.3%). Combining the BUSCO completeness and genome fraction values for the refined RaGOO assembly with its additional contigs results in a genome fraction of 32.815% and a BUSCO completeness value of 81.9%. These values are still lower than the Rebaler assembly values indicating that the Rebaler assembly was the most complete assembly. However, the Rebaler assembly had the highest Ns/100 kb value, which indicated the presence of unresolved gaps following scaffolding.

Table 2.6: Comparison of the Rebaler assembly, refined RaGOO and refined additional RaGOO contig assemblies generated using different reference-guided assembly and polishing pipelines to identify the best *Cannabis sativa* Swazi Gold (SG) genome assembly.

Features	Refined RaGOO assembly	Refined additional RaGOO contigs	Rebaler assembly
Refinement	GapCloser v. 1.12 and Pilon v. 1.23	GapCloser v. 1.12 and Pilon v. 1.23	--
Assembly size (Mb)	438.36	470.40	764.62
N50 value (Mb)	439	499	766
Genome fraction (%)	32.793	0.022	89.87
G+C content (%)	35.15	35.69	34.04
BUSCO completeness (%)	44.6	37.3	97.0
Ns/100 kb value (kb)	1057.73	0.00	2355.59

2.4.9.2. Repeat content of the draft *C. sativa* SG genome assemblies

Repeat content in the assemblies was detected using the RED (Girgis, 2015) and RepeatMasker (Smit *et al.,* 2013). RED repeat analysis revealed a repeat content range of 45.91 to 59.24%, with the highest recorded in the Rebaler assembly. A repeat content of 45.91% was identified within the RaGOO additional contigs (with 52.14% identified in the refined RaGOO assembly), which indicated that a large portion of the contigs removed during RaGOO reference-guided assembly comprised repetitive elements. Overall, the repeat content detected by RED was lower than the expected range (65-78%) for *C. sativa* (Pisupati *et al.,* 2018; Ryu *et al.,* 2024; Shan *et al.,* 2020). This low repeat content could be attributed to limitations associated with RED repeat detection, which include reduced sensitivity for detection of highly diverged transposons and sensitivity to parameters including K-mer length (Girgis, 2015).

RepeatMasker (Smit *et al.,* 2013) demonstrated a repeat content 71.63% within the refined RaGOO assembly, with the additional RaGOO contigs containing 22.6% repeat content. The Rebaler assembly comprised 77.19 % repeat content (Appendix Table 2A).

2.4.9.3. Predicted tRNA and rRNA content

The highest number of tRNAs were predicted within the Rebaler assembly (648 tRNAs) (Figure 2.1), compared to the refined RaGOO assembly (583 tRNAs) and its additional contigs (220 tRNAs).

The predominant tRNAs within the Rebaler assembly included those that carry arginine, leucine and proline; with the least abundant being histidine and tyrosine (Figure 2.1). This was similar to the most abundant tRNAs present in *V. vinifera* and *A. thaliana*, namely those involved in the transfer of arginine and proline (Petrovic *et al.,* 2019; Yao *et al.,* 2020). In contrast, the tRNAs involved in the transfer of leucine, methionine and serine were predominant within the refined RaGOO (with additional contigs) assembly data; with tRNAs that carry histidine and cystine being the least abundant (Figure 2.1). No intron-containing tRNAs were identified within the assembly annotations, which aligns with plant tRNA architecture (Lynch *et al.,* 2025).

More rRNA genes were predicted within the Rebaler assembly than the refined RaGOO (with additional contigs) assembly data (Figure 2.2). The predicted 18S, 28S, and 5S rRNAs, which are essential for ribosome biosynthesis, aligned with the known architecture of eukaryotes (Srivastava & Pal, 2019). Additionally, a total of 278 rRNAs were predicted within the RaGOO additional contigs.

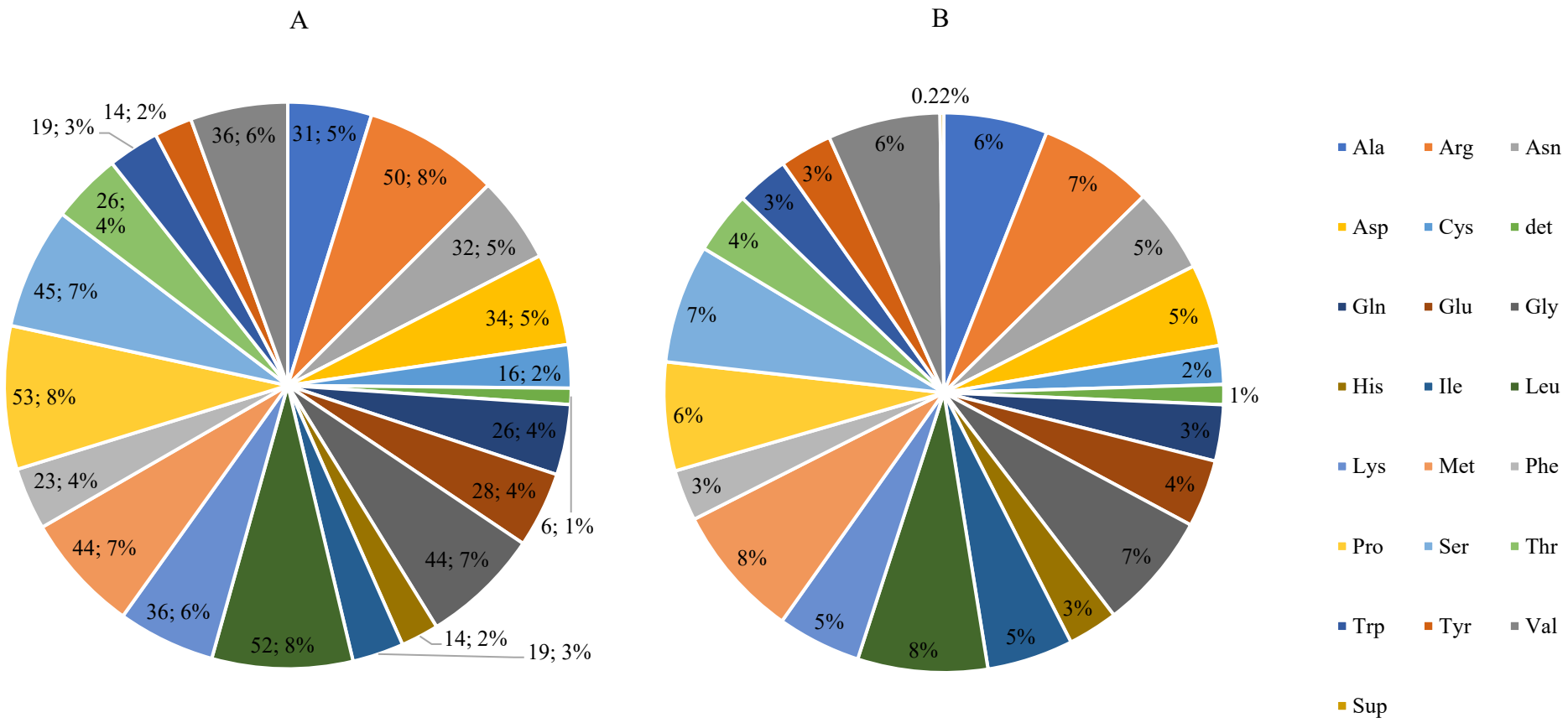


Figure 2.1: Proportion of tRNAs by the amino acid they carry predicted for A) the Rebaler reference-guided and B) the RaGOO reference-guided (with the additional RaGOO contigs) assemblies as predicted using tRNAscan-SE2 v. 2.0. A total of total of 648 tRNAs were predicted within the Rebaler assembly and 1,386 within the RaGOO assembly (with additional contigs).

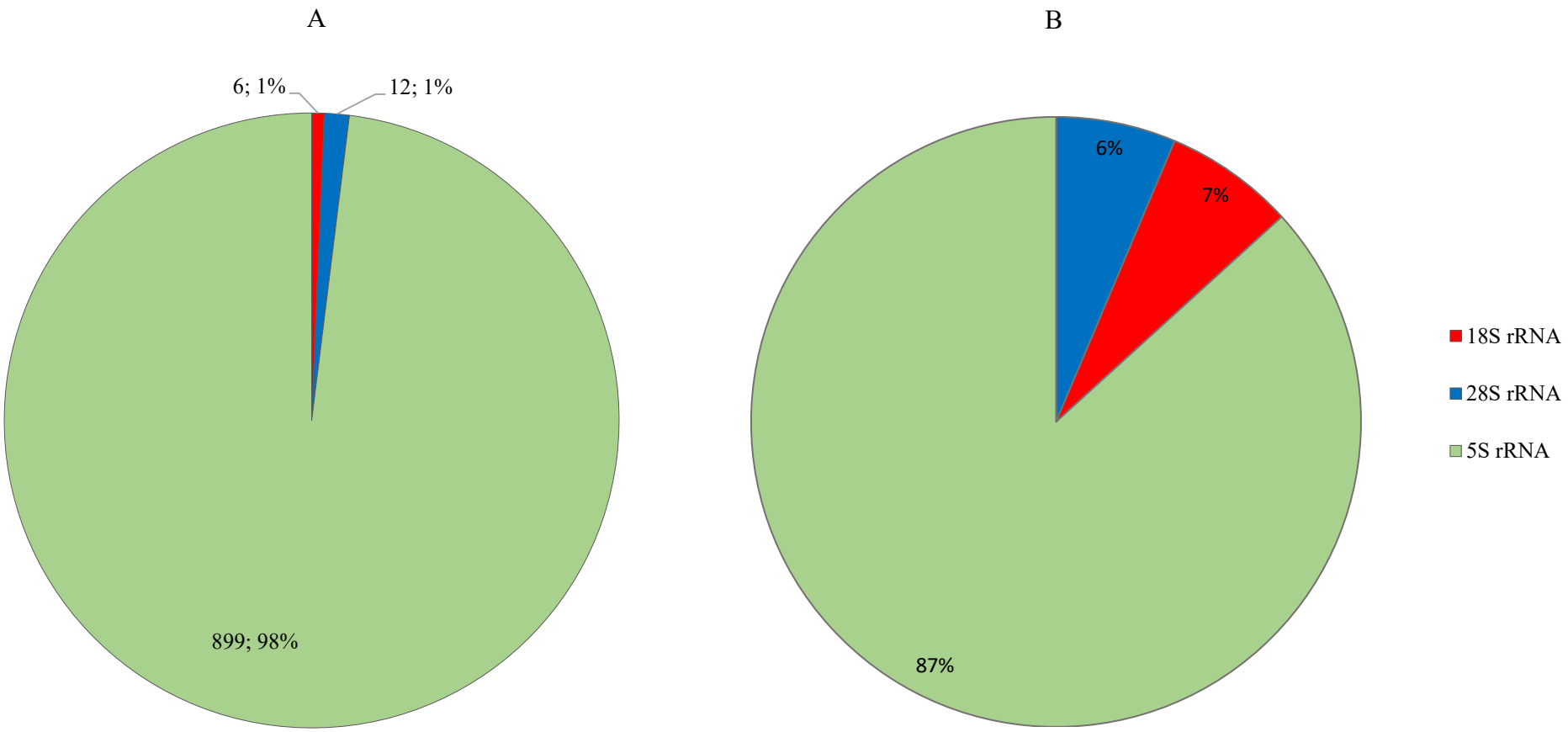


Figure 2.2: Ribosomal RNA (rRNA) composition predicted for A) the Rebaler reference-guided and B) the RaGOO reference-guided (with the additional RaGOO contigs) assemblies as predicted using Barrnap v. 0.9. A total of 907 rRNA genes were predicted within the Rebaler genome, this included 18S rRNA (1%), 28S rRNA (1%) and 5S rRNA (98%). A total of 739 rRNA genes were predicted within the RaGOO (with additional RaGOO contigs) genome, this included 18S rRNA (7%), 28S rRNA (6%) and 5S rRNA (87%).

2.4.9.4. Protein coding sequences

Gene prediction in the Rebaler assembly using AUGUSTUS trained with *V. vinifera* (NCBI RefSeq assembly GCF_030704535.1; Cultivar: Pinot Noir 40024) identified a total of 45,142 genes (Table 2.7). Following structural annotation, eggNOG-mapper functionally annotated 34,950 proteins within the Rebaler reference-guided genome (Huerta-Cepas *et al.,* 2017), while SWISS-PROT functionally annotated 10,206 proteins within the genome (Bairoch & Apweiler, 2000). A total of 43,646 protein-coding genes and 140,786 introns were identified within these genes. Within the refined RaGOO (with additional contigs) assembly data AUGUSTUS identified a total of 98,885 genes and eggNOG-mapper functionally annotated 67,610 proteins, while 453,377 proteins were functionally annotated using SWISS-PROT. This high number of proteins annotated using SWISS-PROT may be attributed to the assignment of more than one annotation per protein due to the multifunctional roles of many proteins (Bairoch & Apweiler, 2000). A total of 97,470 protein-coding genes and 201,570 introns were identified within the refined RaGOO (with additional contigs) genes.

Annotation of the RaGOO assembly and its additional contigs allowed for the identification of more genes, proteins and introns within the RaGOO assembly data compared to the Rebaler assembly. This increase may reflect overlapping predictions (the same genes, proteins or introns) across both of the RaGOO datasets. Notably, more genes, proteins and introns were identified within the additional RaGOO contigs alone than in each of the two assemblies (Table 2.7). This suggests that gene-rich regions of the *C. sativa* SG genome may not have been successfully incorporated within the assembly (linking to the gaps remaining). Similar to the inflated tRNA count within the RaGOO combined assembly data, this increase in predicted annotated features may be attributed to the fragmentation of the contigs leading to gene splitting and assembly artefacts within the assembly data (Denton *et al.,* 2014; Indrischek *et al.,* 2016). Further curation and validation of the assembly will be required to distinguish novel or genuine genes from miss-annotations or redundant predictions.

Table 2.7: Comparison of MOSGA v. 2.1.5 annotation results for the Rebaler assembly, refined RaGOO and refined additional RaGOO contig assemblies for *Cannabis sativa* Swazi Gold.

No. of annotated element	Refined RaGOO reference- guided assembly	Refined additional RaGOO contigs	Rebaler reference-guided assembly
RED repeats (%)	52.14	45.91	59.24
Repeatmasker repeats (%)	71.63	22.06	77.19
tRNAs	469	803	907
rRNAs	583	278	648
Genes	31,687	67,198	45,142
EggNOG proteins	23,981	43,629	34,950
SWISS-PROT proteins	44,820	120,700	10,206
Protein-coding genes	30,381	63,089	43,646
Introns	101,516	100,054	140,786

2.4.10. Selection and evaluation of SG final draft genome

The Rebaler reference-guided genome assembly demonstrated a higher genome fraction (89.87% vs 32.793%), BUSCO completeness value (97% vs 81.9%) and larger genome size (764.62 Mb) than the RaGOO-based assembly. Particularly, the size of the Rebaler assembly is similar to that of the reference genome for *C. sativa* (770.3 Mb). Annotation results of each assembly indicated a higher amount of tRNAs, rRNAs, genes, proteins and introns predicted and identified within the Rebaler assembly compared to the RaGOO assembly alone (without its additional contigs). Consequently, the Rebaler-guided assembly, built on the basis of a MEGAHIT assembly using the paired-end short read data and PacBio data as single-end sequences, was selected for further comparative genome analysis.

2.5. Conclusion

Here we report the first draft genome for *C. sativa* SG assembled using a hybrid *de novo* approach. This approach combined both low coverage long-read (Pacific Bioscience) and high coverage short-read (Illumina NovaSeq 6000) sequencing data. Three short-read assemblers (AbySS, SGA and MEGAHIT) and three hybrid assembly tools (DBG2OLC, MaSurCa and MEGAHIT) were evaluated for optimal parameters and ideal *de novo* assembly. The assembly generated using the MEGAHIT hybrid pipeline outperformed all other evaluated assemblers. Two reference-guided assembly pipelines (RaGOO and Rebaler) were evaluated followed by gap closing with three long-read (LR_Gapcloser, MaSurCa and TGS-GapCloser) and two short-read (ABySS-Sealer and GapCloser) gap closing tools. The gap closed assembly underwent a final round of polishing with Pilon for optimal refinement of the assemblies, producing two refined reference-guided assemblies. These two assemblies (alongside the additional data that was retained post reference-guided assembly with RaGOO) were compared using BUSCO v. 3.0, and structurally and functionally annotated using MOSGA. The two annotated assemblies underwent comparative analysis for selection of the optimal assembly.

The selected final draft genome had a total size of 764.62 Mb, G+C content of 34.04% and a completeness score of 97.0%. A total of 45,142 protein coding genes, 648 tRNAs and 907 rRNAs were predicted within the assembly, while repetitive elements encompassed a predicted 77.19% of the genome. This *C. sativa* SG genome is the first genome assembled for an African *C. sativa* landrace. It provides a key foundation for further comparative genomic analysis among *C. sativa* landraces and cultivars.

2.6. References

Abbas, Q., Wilhelm, M., Kuster, B., Poppenberger, B., Frishman, D. (2024). 'Exploring crop genomes: assembly features, gene prediction accuracy, and implications for proteomics studies', *BMC Genomics*, 25(1), pp. 619. Doi: 10.1186/s12864-024-10521-w.

Arabidopsis Genome Initiative. (2000). 'Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*', *Nature,* 408, pp. 796-815. Doi: 10.1038/35048692

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., Lippman, Z.B., Schatz, M.C. (2019). 'RaGOO: fast and accurate reference-guided scaffolding of draft genomes', *Genome Biology*, 20, pp. 224. Doi: 10.1186/s13059-019-1829-6

Baeza, J.A. (2020). 'Yes, we can use it: a formal test on the accuracy of low-pass whole-genome sequencing for population genomics in the marine invertebrate *Penaeus monodon*', *BMC Genomics*, 21 (729). Doi: 10.1186/s12864-020-07292-8

Bairoch, A., Apweiler, R. (2000). 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Research*, 28(1), pp. 45-48. Doi: 10.1093/nar/28.1.45

Bao, Z., Eddy, S.R. (2002). 'Automated *de novo* identification of repeat sequence families in sequenced genomes', *Genome Research*, 12(8), pp. 1269-1276. Doi:10.1101/gr.88502.

Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M., Phillippy, A.M. (2015). 'Assembling large genomes with single-molecule sequencing and locality-sensitive hashing', *Nature Biotechnology*, 33(6), pp. 623-630. Doi: 10.1038/nbt.3238

Bertrand, D., Garrison, E., Marth, G. (2019). 'OPERA-MS: Hybrid genome assembly with long and short reads', *Bioinformatics*, 35(20), pp. 4044-4046. Doi: 10.1093/bioinformatics/btz233

Benson, G. (1999). 'Tandem repeats finder: a program to analyse DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–-80. Doi:10.1093/nar/27.2.573.

Boetzer, M., Pirovano, W. (2014). 'SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information', *BMC Bioinformatics*, 15(1), pp. 211. Doi: 10.1186/1471-2105-15-211

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.C., Corbeil, J., Del Fabbro, C., Docking, T.R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N.A., Ganapathy, G., Gibbs, R.A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J.B., Ho, I.Y., Howard, J., Hunt, M., Jackman, S.D., Jaffe, D.B., Jarvis, E.D., Jiang, H., Kazakov, S., Kersey, P.J., Kitzman, J.O., Knight, J.R., Koren, S., Lam, T.W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M.D., Maillet, N., Melnikov, S., Naquin, D., Ning,

Z., Otto, T.D., Paten, B., Paulo, O.S., Phillippy, A.M., Pina-Martin, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F.J., Richards, S., Rokhsar, D.S., Ruby, J.G., Scalabrin, S., Schatz, M.C., Schwartz, D.C., Sergushichev, A., Sharpe, T., Shaw, T.I., Shendure, J., Shi, Y., Simpson, J.T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B.M., Wang, J., Worley, K.C., Yin, S., Yiu, S.M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., Zhu, J., Zimin, A.V., and Korf, I.F. (2013). 'Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species', *GigaScience*, 2(1), pp. 10. Doi: 10.1186/2047-217X-2-10

Casano, S., Grassi, G., Martini, V., Michelozzi, M. (2010). 'Variations in Terpene profile of different strains of *Cannabis sativa* L.', *ISHS Acta Horticulture*, 925. Doi: 10.17660/ActaHortic.2011.925.15

Chen, Z., Erickson, D.L., Meng, J. (2020). 'Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford nanopore sequencing', *BMC Genomics*, 21, pp. 631. Doi: 10.1186/s12864-020-07041-8

Chikhi, R. and Rizk, G., 2013. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms for Molecular Biology*, 8, pp. 22. Doi: 10.1186/1748-7188-8-22

Cook, D., Bolt, B.J., Chang, H.Y., El-Gebali, S., Fraser, M., Kulesha, E., McLaren, W.M., Maheswari, U., Nuhn, M., Papatheodorou, I., Papasotiriou, D., Pignatelli, M., Riat, H.S., Richardson, L., Ruffier, M., Schuster, M., Spooner, W., Yates, A., Zadissa, A., Birney, E., Harrow, J. (2017). 'The European Bioinformatics Institute in 2016: Data growth and integration', *Nucleic Acids Research*, 45(D1), pp. D18-D25. Doi: 10.1093/nar/gkw1012

Cullings, K.W. (1992). 'Design and testing of a plant-specific PCR primer for ecological and evolutionary studies', *Molecular Ecology*, 1, pp. 233-240. Doi: 10.1111/j.1365-294X.1992.tb00182.x

Cuperus, J.T., Queitsch, C. (2020). 'Editorial overview: Technology development as a driver of biological discovery', *Current Opinion in Plant Biology*, 54, pp. A1-A4. Doi: 10.1016/j.pbi.2020.06.001

Crampton, H. (2015). 'Dagga. A short History (then, now &just now)', *South Africa: Jacana*. Available at: https://books.google.co.za/books/about/Dagga.html?id=XTjqsgEACAAJ&redir_esc=y (Accessed 16/03/2024)

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H. (2021). 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2), giab008. Doi:10.1093/gigascience/giab008.

Darwin Tree of Life Project Consortium. (2022). 'Sequence locally, think globally: The Darwin Tree of Life Project', *Proceedings of the National Academy of Sciences*, 119(4), pp. e2115642118. Doi: 10.1073/pnas.2115642118

D'Auria, S. (2017). 'The Swiss Institute of Bioinformatics: A bioinformatics service provider for life science research', *Nucleic Acids Research*, 45(D1), pp. D12-D17. Doi: 10.1093/nar/gkw995

Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C., Hahn, M.W., (2014). 'Extensive error in the number of genes inferred from draft genome assemblies', *PLoS Computational Biology*, 10(12), pp. e1003998. Doi: 10.1371/journal.pcbi.1003998

Doyle, J.J., Doyle, J.L. (1987). 'A rapid DNA isolation procedure for small quantities of fresh leaf tissue', Phytochemistry Bulletin, 19, pp.11-15. Available at: https://webpages.charlotte.edu/~jweller2/pages/BINF8350f2011/BINF8350_Readings/Doyle_plantDNAextractCTAB_1987.pdf (Accessed 12/08/2024)

Dumschott, K., Schmidt, M.H., Chawla, H.S., Snowdon, R., Usadel, B. (2020). 'Oxford Nanopore sequencing: new opportunities for plant genomics?', *Journal of Experimental Botany*, 71 (18), pp. 5313-5322. Doi: 10.1093/jxb/eraa263

Ewing, B., Green, P. (1998). 'Base-calling of automated sequencer traces using phred. II. Error probabilities', *Genome Research*, 8(3), pp. 186-194. Available at: https://doi.org/10.1101/gr.8.3.186 (Accessed 20/09/2025)

Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F. (2020) 'RepeatModeler2 for automated genomic discovery of transposable element families', *Proceedings of the National Academy of Sciences*, 117(17), pp. 9451-9457. Doi:10.1073/pnas.1921046117.

Garcia, S., Crhák Khaitová, L. & Kovařík, A. (2012). 'Expression of 5 S rRNA genes linked to 35 S rDNA in plants, their epigenetic modification and regulatory element divergence', *BMC Plant Biology*, 12 (95). Doi: 10.1186/1471-2229-12-95

Gao, S., Wang, S., Xie, S., Xu, X., Zhang, J., Pei, L., Yu, Y., Yang, W. and Zhang, Y. (2020) 'A high-quality reference genome of wild *Cannabis sativa*', *Horticulture Research*, 7, pp. 73. Doi:10.1038/s41438-020-0295-3.

Girgis, H.Z. (2015). 'Red: an intelligent, rapid, accurate tool for detecting repeats *de novo* on the genomic scale', *BMC Bioinformatics*, 16, pp. 227. Doi: 10.1186/s12859-015-0654-5

Glasel, J.A. (1995). 'Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios', *Biotechniques*, 18(1), pp. 62-3. Available at: https://pubmed.ncbi.nlm.nih.gov/7702855/ (Accessed 18/09/2024)

Guigó R. (2023). ‘Genome annotation: From human genetics to biodiversity genomics’*, Cell Genomics*, 3(8). Doi: 10.1016/j.xgen.2023.100375.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G. (2013). ‘QUAST: quality assessment tool for genome assemblies’, *Bioinformatics*, 29(8), pp. 1072-1075. Doi: 10.1093/bioinformatics/btt086

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., White, O. (2003). ‘Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies’, *Nucleic Acids Research*, 31(19), pp. 5654-5666. Doi: 10.1093/nar/gkg770

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P. (2017). ‘Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper’, *Molecular Biology and Evolution*, 34(8), pp. 2115-2122. Doi: 10.1093/molbev/msx148

Indrischek, H., Wieseke, N., Stadler, P.F., Prohaska, S.J., Bernt, M., (2016). ‘The paralog-to-contig assignment problem: high quality gene models from fragmented assemblies, *Algorithms for Molecular Biology*, 11(1), pp. 1. Doi: 10.1186/s13015-016-0063-y

Illumina (2024). ‘Sample Input for Illumina DNA Prep (M) Tagmentation Library Preparation Kit – FAQ’, *Illumina Knowledge Base*. Available at: https://knowledge.illumina.com/library-preparation/dna-library-prep/library-preparation-dna-library-prep-faq-list/000001262 (Accessed 09/09/2025)

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., S. Austin Hammond, Jahesh, G., Khan, H., Coombe, L., Warren, R.L., Birol, I. (2017). ‘ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter’, *Genome Research*, 27(5), pp. 768-777. Doi:10.1101/gr.214346.116

Kong, N., Ng, W., Thao, K., Agulto, R., Weis, A., Spittle Kim, K., Korlach, J., Hickey, L., Kelly, L., Lappin, S., & Weimer, B. C. (2017). ‘Automation of PacBio SMRTbell NGS library preparation for bacterial genome sequencing’, *Standards in Genomic Sciences*, 12, pp. 27. Doi: 10.1186/s40793-017-0239-1

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M. (2017). ‘Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation’, *Genome Research*, 27(5), pp. 722-736. Doi: 10.1101/gr.215087.116

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J.,

Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., DePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Diop, S.I., Elliott, A., Ene, C.O., Eom, A.H., Feng, T., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gitzendanner, M.A., Graham, S.W., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Juniper, B., Kates, H.R., Kessler, M., Kidner, C.A., Kohchi, T., Kovar, L., Krause, K., Krupinska, K., Kuo, L.-Y., Kuo, Y.-M., Labadie, K., Lalonde, M., Lam, V.K.Y., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Li, J., Li, Q., Li, Z., Liu, B., Liu, C., Liu, J., Liu, X., Llorens, C., Luan, S., Matasci, N., McAbee, J.M., McKain, M.R., McLellan, T., McNeal, J.R., McPherson, M.A., McSteen, P., Mendez, L., Mikulic, P., Ming, R., Mower, J.P., Murch, S.J., Nelson, D.R., Nevill, P.G., Newell, C.A., Newton, A.E., Nge, F.J., Nguyen, T., Noutsos, C., Novikova, P.Y., Oldenkott, B., O'Neill, C.M., O'Reilly, J.E., Otto, S.P., Owens, G.L., Petersen, G., Pires, J.C., Poncet, V., Proctor, V., Radhakrishnan, D., Ramesh, S., Rest, J.S., Rexroad, C., Rieseberg, L.H., Rothfels, C.J., Sage, R.F., Sakai, A.K., Salmon, A., Sander, R., Sanguinet, K.A., Schilling, E.E., Schmutz, J., Schulz, M.H., Schwartz, J.A., Scott, A.D., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.R., Smithson, A., Stewart, C.N. Jr., Stinchcombe, J.R., Surek, B., Thomas, P., Villarreal, J.C., Waller, D.M., Warnow, T., Weng, J.-K., Wicke, S., Wu, X., Wullschleger, S.D., Yang, Y., Zhu, D., Zhuang, C., Zinkgraf, M., Wong, G.K.-S., Matasci, N. (2019). 'One thousand plant transcriptomes and the phylogenomics of green plants', *Nature*, 574(7780), pp. 679-685. Doi: 10.1038/s41586-019-1693-2

Lewin H.A., Richards S., Lieberman Aiden E., Allende M.L., Archibald J.M., Bálint M., Barker K.B., Baumgartner B., Belov K., Bertorelle G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., Durbn, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Thomas, M., Gilbert, P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O'Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C., Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D. and Zhang, G. (2022). 'The Earth BioGenome Project: Starting the clock', *Proceedings of the National Academy of Sciences*, 119, article 4. Doi: 10.1073/pnas.2115635118

Li, H., Durbin, R. (2009). 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*, 25(14), pp.1754-1760. Doi:10.1093/bioinformatics/btp324

Li, D., Liu, C.M., Luo, R., Sadakane, K., Lam, T.W. (2015). 'MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph', *Bioinformatics*, 31(10), pp. 1674-1676. Doi: 10.1093/bioinformatics/btv033

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., Lam, T.-W. (2016). 'MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler driven by Advanced Methodologies and Community Practices', *Methods*, 102, pp. 3-11. Doi: 10.1016/j.ymeth.2016.02.020.

Li, H. (2018). 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*, 34(18), pp. 3094-3100. Doi: 10.1093/bioinformatics/bty191

Li, H. (2018). 'Seqtk (Version 1.3-r106)' [Software]. Available at: https://github.com/lh3/seqtk (Accessed 21/09/2025).

Li, H., Schatz, M.C. (2025). 'Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions', *Quantitative Plant Biology*, 6. Doi: 10.1017/qpb.2025.18

Lowe, T. M. & Chan, P. P. (2016). 'tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes', *Nucleic Acids Res*earch, 44, pp. W54-W57. Doi: 10.1093/nar/gkw413

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.W., Wang, J. (2012). 'SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler', *GigaScience*, 1(1), p. 18. Doi: 10.1186/2047-217X-1-18

Lynch, R.C., Padgitt-Cobb, L.K., Garfinkel, A.R., Knaus, B.J., Hartwick, N.T., Allsing, N., Aylward, A., Bentz, P.C., Carey, S.B., Mamerto, A., Kitony, J.K., Colt, K., Murray, E.R., Duong, T., Chen, H.I., Trippe, A., Harkess, A., Crawford, S., Vining, K., Michael, T.P. (2025). 'Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome', *Nature,* 643(1001–1010), pp. 1001-1010. Doi: 10.1038/s41586-025-09065-0

MacManes, M.D. (2014) 'On the optimal trimming of high-throughput mRNA sequence data', *Frontiers in Genetics*, 5, pp. 13. Doi:10.3389/fgene.2014.00013.

Mardis, E.R. (2008). 'Next-generation DNA sequencing methods'. *Annual Review of Genomics and Human Genetics*, 9, pp. 387-402. Available at: https://doi.org/10.1146/annurev.genom.9.081307.164359 (Accessed 20/09/2025)

Marks, R.A., Hotaling, S., Frandsen, P.B., VanBuren, R. (2021). 'Representation and participation across 20 years of plant genome sequencing". *Nature Plants*, 7(12), pp. 1571-1578. Available at: https://doi.org/10.1038/s41477-021-01031-8 (Accessed 20/09/2025)

Martin, M. (2011). 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp.10-12. Doi: 10.14806/ej.17.1.200

Martin, R., Hackl, T., Hattab, G., Fischer, M.G., Heider, D. (2020). 'MOSGA: Modular open-source genome annotator', *Bioinformatics, 3*6(22–23), pp. 5514-5515. Doi: 10.1093/bioinformatics/btaa1003

Michael, T.P., VanBuren, R. (2020). 'Building near-complete plant genomes', *Current Opinion in Plant Biology*, 54, pp. 26-33. Doi:10.1016/j.pbi.2019.12.009

Ndlangamandla, V.V., Salawu-Rotimi, A., Bushula-Njah, V.S., Hlongwane, N.L., Sibandze, G.F., Gebashe, F.C. & Mchunu, N.P. (2024). 'Finally freed – *Cannabis* in South Africa: A review contextualised within global history, diversity, and chemical profiles', *Plants (Basel)*, 13(19), pp. 2695. Doi: 10.3390/plants13192695

Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., Koren, S. (2020). 'HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads', *Genome Research*, 30(9), pp. 1291-1305. Doi: 10.1101/gr.263566.120

PacBio (2021). 'Application note — Considerations for using the low and ultra-low DNA input workflows for whole genome sequencing', Available at: https://www.pacb.com/ (Accessed 09/09/2025)

Pacific Biosciences of California Inc. (2021). 'Procedure & Checklist – Preparing HiFi SMRTbell Libraries using the SMRTbell Express Template Prep Kit 2.0', Version 05, August 2021). Menlo Park, CA: Pacific Biosciences of California Inc. Available at: https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf (Accessed 8/09/2025).

Paulino, D., Warren, R.L., Vandervalk, B.P., Raymond, A., Jackman, S.D., Birol, I. (2015). 'Sealer: a scalable gap-closing application for finishing draft genomes', *BMC Bioinformatics*, 16, pp. 230. Doi: 10.1186/s12859-015-0663-4

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L. (2012). 'IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth', *Bioinformatics*, 28(11), pp. 1420-1428. Doi: 10.1093/bioinformatics/bts174

Petrovic, G., Aleixandre-Tudo, J.L., Buica, A. (2019). 'Grape must profiling and cultivar discrimination based on amino acid composition and general discriminant analysis with best subset', *South African Journal of Enology and Viticulture*, 40, article 2. Doi: 10.21548/40-2-3373

Pillay, M., Kenny, S.T. (2006). 'Structural organization of the nuclear ribosomal RNA genes in *Cannabis* and *Humulus* (*Cannabaceae*)', *Plant Systems and Evolution,* 258, pp. 97-104. Doi: 10.1007/s00606-005-0396-3

Pisupati, R., Vergara, D., Kane, N.C. (2018). 'Diversity and evolution of the repetitive genomic content in *Cannabis sativa*', *BMC Genomics*, 19, pp. 156. Doi:10.1186/s12864-018-4494-3

Price, A.L., Jones, N.C., Pevzner, P.A. (2005). '*De novo* identification of repeat families in large genomes', *Bioinformatics*, 21(Suppl 1), pp. i351-i358. Doi:10.1093/bioinformatics/bti1018.

Quinlan, A.R., Hall, I.M. (2010). 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841-842. Doi:10.1093/bioinformatics/btq033.

Rhie, A., McCarthy, S.A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B.J., Chaisson, M.J.P., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., Haase, B., Mountcastle, J., Winkler, S., Paez, S., Howard, J., Stitzer, M.C., Armstrong, J., Balkema-Buschmann, A., Bruno, O., Calero-Cáceres, W., Camacho-Sanchez, M., Clark, K., Cornelli, S., Danis, D., Daza-Cajigal, V., de Clare, M., Edgson, R., Fiddes, I.T., Forcina, G., Gut, I., Haggart, D., Hansen, N.F., Heaton, H., Hillier, L.W., Howie, J.M., Hunkapiller, M.W., Jain, C., Johnson, J., Khan, Z., Korlach, J., Kremitzki, M., Lam, J., Lawniczak, M.K.N., Lister, C.E., Livermore, S., Lozano-Fernandez, J., Marschall, T., Matthews, L., McLay, K., Mead, D., Meier, J., Mellis, C., Mills, R.E., Morisson, J., Muffato, M., Noll, A., Palmer, J., Peck, M., Phillippy, A.M., Rogaev, E., Rosenfeld, J.A., Rye, A., Salazar-Torres, J.J., Schmitt, A.D., Schneider, V.A., Sello, S., Siol, M., Smetanin, D., Smith, C., Stainton, D., Stevens, C., Sutton, G., Thomas, M., Tracey, A., Tricomi, F.F., Uliano-Silva, M., Underwood, C.J., Wright, J.C., Young, E.R., Yudkin, G., Zeng, J., Zhang, G., Zhang, H., Jarvis, E.D., Warren, W.C., Howe, K., Flicek, P., Koepfli, K.P., Paten, B., Haussler, D., Korlach, J., Li, H. and Phillippy, A.M. (2021). 'Towards complete and error-free genome assemblies of all vertebrate species', *Nature*, 592(7856), pp.737-746. Doi:10.1038/s41586-021-03451-0

Ruan, J., Li, H. (2020). 'Fast and accurate long-read assembly with wtdbg2', *Nature Methods*, 17(2), pp. 155-158. Doi: 10.1038/s41592-019-0669-3

Ryu, BR., Gim, GJ., Shin, YR., Kang, M.J., Kwon, T.H., Lim, Y.S., Parl, S.H., Lim, J.D. (2024). 'Chromosome-level haploid assembly of *Cannabis sativa* L. cv. Pink Pepper', *Scientific Data*, 11, pp. 1442. Doi: 10.1038/s41597-024-04288-8

Sambrook, J., Russell, D.W. (2001). 'Molecular Cloning: A Laboratory Manual 3rd edition', Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. Available at https://www.cshlpress.com/pdf/sample/2013/MC4/MC4FM.pdf (Accessed 06/09/2025)

Seemann, T. (2018). 'Barrnap 0.9: rapid ribosomal RNA prediction [Internet]' *GitHub repository*, Available at https://github.com/tseemann/barrnap (Accessed 16/08/2025)

Schatz, M.C., Langmead, B., Salzberg, S.L. (2010). 'Cloud computing and the DNA data race.', *Nature Biotechnology*, 28(7), pp. 691-693. Doi:10.1038/nbt0710-691

Shrestha, R.K., Lubinsky, B., Bansode, V.B., Moinz, M.B.J., McCormack, G.P., Travers, S.A. (2014) 'QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform', *BMC Bioinformatics*, 15(1), article 33. Doi:10.1186/1471-2105-15-33.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, İ. (2012). 'A by-product of *de novo* sequencing: Assembly on a string graph', *Genome Research*, 22(2), pp. 290-300. Doi: 10.1101/gr.126369.111

Simpson, J.T., Durbin, R. (2012). 'Efficient *de novo* assembly of large genomes using compressed data structures', *Genome Research*, 22(3), pp. 549-56. Doi: 10.1101/gr.126953.111

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., Birol, I. (2009). 'ABySS: a parallel assembler for short read sequence data', *Genome Research*, 19(6), pp. 1117-1123. Doi:10.1101/gr.089532.108

Smit, A. F. A., Hubley, R., Green, P. (2013). *RepeatMasker Open-4.0*. Available at: http://www.repeatmasker.org/RMDownload.html (Accessed 08/04/2024)

Srivastava, P., & P. B. Pal, (2019). 'Ribosome biogenesis in plants: From functional 45S rRNA genes to ribosome subunits', *Frontiers in Plant Science*, 10, pp. 1198. Doi: 10.1105/tpc.18.00874

Schwacke, R., Bolger, M.E., Usadel, B. (2025). 'PubPlant – a continuously updated online resource for sequenced and published plant genomes'. *Frontiers in Plant Science*, 16, pp. 1603547. Available at: https://doi.org/10.3389/fpls.2025.1603547 (Accessed 20/09/2025)

Van der Walt, A.J., van Goethem, M.W., Ramond, J.-B., Makhalanyane, T.P., Reva, O. and Cowan, D.A. (2017) 'Assembling metagenomes, one community at a time', *BMC Genomics*, 18, p. 521. Doi: 10.1186/s12864-017-3918-9.

Vaser, R., Sovic, I., Nagarajan, N., Sikic, M. (2017). 'Racon: Fast and accurate consensus module for raw *de novo* genome assembly of long uncorrected reads', *Genome Research*, 27(5), pp.677-685. Doi: 10.1101/gr.214270.116

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). 'Pilon: An integrated tool for comprehensive

microbial variant detection and genome assembly improvement', *PLOS ONE*, 9, article 11. Doi: 10.1371/journal.pone.0112963

Wick, R.R., Judd, L.M. and Holt, K.E. (2019). 'Performance of neural network base calling tools for Oxford Nanopore sequencing', *Genome Biology*, 20, pp. 129. Doi: 10.1186/s13059-019-1727-y

Xu, G.-C., Xu, T.-J., Zhu, R., Zhang, Y., Li, S.-Q., Wang, H.-W., Li, J.-T. (2018). 'LR_Gapcloser: a tiling path–based gap closer that uses long reads to complete genome assembly', *GigaScience*, 8, article 1. Doi: 10.1093/gigascience/giy157

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., Zhang, Y. (2020). 'TGS GapCloser: A fast and accurate gap closer for large genomes with low coverage of error prone long reads', *GigaScience*, 9, article 9. Doi: 10.1093/gigascience/giaa094

Yao, X., Nie, J., Bai, R., Sui, X. (2020). 'Amino acid transporters in plants: Identification and function', *Plants (Basel*), 9(8), pp. 972. doi: 10.3390/plants9080972.

Ye, C., Hill, C.M., Wu, S., Ruan, J., Ma, Z.S. (2016). 'DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third-generation sequencing technologies', *Scientific Reports*, 6. Doi: 10.1038/srep31900

Zerbino, D.R., Birney, E. (2008). 'Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs', *Genome Research*, 18(5), pp. 821-829. Doi: 10.1101/gr.074492.107

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A. (2013). 'The MaSuRCA genome assembler', *Bioinformatics*, 29(21), pp. 2669-2677. Doi: 10.1093/bioinformatics/btt476

Zimin, A.V., Salzberg, S.L. (2022). 'The SAMBA tool uses long reads to improve the contiguity of genome assemblies', *PLoS Computational Biology*, 18(2), pp. e1009860. Doi: 10.1371/journal.pcbi.1009860

Zhao, Y., Zhang, J., Xu, T., Wang, W. (2023). 'Long-read sequencing (method of the year 2022): The way forward for plant omics research', *Molecular Plant*, 16(6), pp. 965-967. Doi: 10.1016/j.molp.2023.04.002

Chapter Three: Comparative genomic analysis of *Cannabis sativa* Swazi Gold

3.1. Abstract

Cannabis sativa Swazi Gold (SG) is regarded as a Type I tetrahydrocannabinol (THC) dominant chemotype (drug-type) plant originating from the Kingdom of Eswatini. Here the genome of this landrace was compared against the publicly available genomes of fourteen *C. sativa* cultivars. Core genome phylogeny demonstrated mixed chemotype clustering across the fifteen *C. sativa* genomes compared in this study. BLAST-based identification and alignment of genes involved in cannabinoid biosynthesis (*ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS*) revealed extensive paralogue divergence and copy number variation (CNV) amongst the cultivars and landraces. Biosynthetic gene cluster (BGC) profiling identified 31 distinct BGCs within the *C. sativa* SG genome, and a total 501 BGCs across all fifteen annotated *C. sativa* genomes. The most abundant BGCs included saccharides, alkaloids and terpenes. These findings suggest chemotype status cannot simply be inferred by the assessing the genome of a cultivar for presence and copy number of a specific cannabinoid biosynthetic gene but will instead have to be validated through more complex chemical profiling methods.

3.2. Introduction

Comparative genomic analysis of multiple plant genomes can provide insight into their genetic architecture, structural and functional characteristics, their evolutionary histories and potential applications (Andre *et al.,* 2016; van Bakel *et al.,* 2011). In the case of *Cannabis sativa*, these analyses are crucial to evaluate genomic divergence due to domestication and artificial selection and infer key genetic elements involved in phenotypic variation and adaptation, such as variation in cannabinoid biosynthesis for protection against herbivores and microbial pathogens (Kojoma *et al.,* 2006; van Bakel *et al.,* 2011).

Comparative analysis of cannabinoid biosynthetic genes has identified multiple single nucleotide polymorphisms (SNPs) within tetrahydrocannabinolic acid synthase (THCAS), cannabidiolic acid synthase (CBDAS) and cannabichromenic acid synthase (CBCAS); which may alter the cannabinoid profile of the plant (Laverty *et al.,* 2019, Rotherham & Harbison, 2010, Taura *et al.,* 2007). The selection and upregulation of Δ^9-THCAS over CBDAS within a cultivar leads to elevated THC production characteristic of drug-type chemotypes (de Meijer *et al.,* 2009; van Bakel *et al.,* 2011). By contrast, fibre-type chemotypes favour CBDAS activity (Morimoto *et al.,* 1998; van Bakel *et al.,* 2011). *Cannabis sativa* exhibits heterozygosity rates

ranging from 0.18-0.26%, with drug- and fibre-type strains being separated by SNPs (occurrence of SNPs = 0.64%) (van Bakel *et al.,* 2011). Major SNPs located within *THCAS* and *CBDAS* separate chemotypes I, II and III (Pacifico *et al.,* 2006).

A comprehensive comparison of *C. sativa* genomes with different chemical profiles (chemotypes) would provide valuable insight into conserved gene families, regulatory elements and sequence variations between landraces and cultivars across long periods of evolutionary time (Stott & Bobay, 2020; You, 2023). The identification of BGCs within the fifteen *C. sativa* genomes may provide insight into the chemotypes, aroma and flavour profiles (Kautsar *et al.,* 2017; Zager *et al.,* 2019). Here we have performed foundational comparative genomic analyses of African landrace *C. sativa* Swazi Gold (SG) and the high-quality genome assemblies of fourteen *C. sativa* cultivars and landraces. It is envisaged that this research will contribute towards a deeper understanding of African landraces' cannabinoid production profile, with the potential end goal of therapeutic application of these phytochemicals.

3.3. Methods and Materials

3.3.1. Structural and functional annotation of available genomes

The assembled genome of *C. sativa* SG was compared to fourteen other *C. sativa* genomes, thirteen of which are publicly available on the NCBI database (Table 3.1), and one in-house draft genome assembly of the South African landrace, Durban Poison (DP). All genomes were structurally and functionally annotated using MOSGA v. 2.1.5 (Martin *et al.,* 2020) as outlined in Chapter 2 (section 2.3.9). Repeat content was determined using the REpeat Detector (RED) pipeline (Girgis, 2015), as well as RepeatMasker v. 4.2.0 (Smit *et al.,* 2013), as outlined in Chapter 2 (section 2.3.9). This was followed by functional annotation and classification of genes and proteins using eggNOG-mapper v. 2.1.13 and SWISS-PROT (Bairoch & Apweiler, 2000; Huerta-Cepas *et al.,* 2017), as outlined in Chapter 2 (section 2.3.9). Additionally, these genomes underwent comparison against the eudicots_odb10 lineage dataset (n = 2,326 BUSCOs) using BUSCO v. 3.0 (Bradnam *et al.,* 2013; Waterhouse *et al.,* 2018) to assess genome completeness.

Table 3.1: Overview of fifteen comparator *Cannabis sativa* genomes with chemotype and geographic location data. Type I: THC dominant (drug-type); Type II: THC:CBD (intermediate); Type III: (CBD dominant (fibre-type).

Cultivar / Landrace	GenBank Assembly Accession #	Plant Chemotype	Geographic origin of the plant
Pink pepper (ref)	GCA_029168945.1	Type III	South Korea
Abacus (Csat)	GCA_025232715.1	Type III	United States of America (USA)
Cannbio-2	GCA_016165845.1	Type II	Australia
Cherry Pie	GCA_049724855.1	Type I	California USA
cs10	GCF_900626175.2	Type III	Colorado, USA
Durban Poison (DP)	In house assembly (unpublished)	Type I	Durban, South Africa
Finola	GCA_003417725.2	Type III	Finland
JL	GCA_013030365.1	Unclassified	Jilong, China
Jamaican Lion Dash (JL Dash)	GCA_003660325.2	Type II	Massachusetts, USA
Jamaican Lion ^4 (JL5)	GCA_016415525.1	Type II	Massachusetts, USA
Jamaican Lion ^4 (JL Mother)	GCA_012923435.1	Type II	Massachusetts, USA
Jamaican Lion ^4 (JL Father)	GCA_013030025.1	Type II	Massachusetts, USA
Punto Rojo	GCA_049725155.1	Type I	Colombia
Purple Kush	GCA_000230575.5	Type I	California USA

3.3.2. Single-copy orthologue identification and core genome phylogeny construction

A core genome phylogeny was generated to evaluate the evolutionary relationships between the *C. sativa* cultivars and landraces. The protein datasets for each genome from MOSGA v.2.1.5 were compared using OrthoFinder v. 3.1.0 (Emms & Kelly, 2019) to identify single-copy orthologues (SCOs). Single-copy orthologues refer to proteins that are encoded on all compared genomes and that have not undergone duplication (Emms & Kelly, 2019). The identified SCOs were aligned individually using the mCOFFEE implementation of T-Coffee v. 13.41.0.28bdc39 (Notredame *et al.,* 2000) and concatenated using an in-house perl script. The concatenated alignment was treated with Gblocks v. 0.91b (Castresana, 2000) to remove poorly aligned blocks. The curated concatenated alignment was used to construct a maximum likelihood (ML) phylogeny with IQ-Tree3 (Wong *et al.,* 2025). Branch support of the ML phylogeny was inferred with ultrafast bootstrapping (n = 1,000 replicates) using UFBoot2 (Hoang *et al.,* 2018).

3.3.3. BLAST analysis of cannabinoid biosynthetic genes in *C. sativa*

The reference protein sequences of all the enzymes involved in the biosynthesis of cannabinoids were retrieved from the NCBI database (Table 3.2). These sequences were compared against the genome sequences of each *C. sativa* cultivar/landrace by local tBlastn analysis implemented in BioEdit v. 7.2.5 (Hall, 1999). The presence of these specific cannabinoid biosynthesis genes within the genomes would suggest that all compared cultivars and landraces are capable of cannabinoid production, however, future chemical profile analysis is required to confirm pathway functionality and cannabinoid production.

The tBLASTn output was examined to identify paralogues and determine amino acid sequence identity across the *C. sativa* genomes. Paralogues refer to genes in an organism which evolved from gene duplication; paralogous proteins experience evolutionary pressure that often leads to altered or novel specificities and functions (Mirny & Gelfand, 2002). Multiple sequence alignments of the identified paralogues were conducted using MUSCLE in MEGA12 (Kumar *et al.,* 2018). These aligned sequences were used for the construction of neighbour-joining phylogenetic trees in MEGA12. Separate phylogenetic trees were generated for each protein sequence to compare the identified paralogues across the fifteen *C. sativa* genomes.

Additionally, the MUSCLE aligned sequences for the *C. sativa* SG paralogues were used to generate a sequence identity matrix using BioEdit v. 7.2.5.

Table 3.2: Overview of *Cannabis sativa* cannabinoid biosynthetic proteins retrieved from the NCBI database.

Protein	**Enzyme**	**NCBI GenBank Accession #**
ACCase	Acetyl-CoA carboxylase	XM_030629499
AP	Aromatic prenyltransferase	MN551226
CBDAS	Cannabidiolic acid synthase	MW382907.1
CBCAS	Cannabichromenic acid synthase	XM_061118802.1
DXS1	1-Deoxy-D-xylulose-5-phosphate synthase	XM_061104475.1
DXS2	1-Deoxy-D-xylulose-5-phosphate synthase	KY014577.1
HCAS	Hexanoyl-CoA synthetase	XM_030637946.2
GPPS1	Geranyl pyrophosphate synthase	KY014573.1
GPPS2	Geranyl pyrophosphate synthase	KY014574.1
GOT	Geranylpyrophosphate olivetolate geranyltransferase	XM_061106831.1
OAC	Olivetolic acid cyclase	NM_001397939.1
OLS / TKS	Olivetol synthase / Tetraketide synthase	AB164375.1
THCAS	Tetrahydrocannabinolic acid synthase	JQ437488.1

3.3.4. Biosynthetic gene cluster (BGC) identification

Biosynthetic gene clusters (BGCs) are dense clusters of genes encoding the biosynthetic pathway of specialized metabolites, such as tri-terpenes in *Arabidopsis* species (Huang *et al.,* 2019) and morphine in opium poppy (Guo *et al.,* 2018; Jensen, 2016; Satria *et al.,* 2017). PlantiSMASH (Kautsar *et al.,* 2017) set to default parameters was used for the identification of candidate BGCs within the *C. sativa* genomes. The software utilises Hidden Markov Models (pHMMs) of known enzyme families in plant biosynthetic pathways and CD-HIT clustering of protein sequences belonging to the same family. This allows for the identification and grouping of functionally related genes into BGCs.

3.4. Results and Discussion

3.4.1. Comparison of the annotated *C. sativa* genomes

3.4.1.1. Genomic metrics of *C. sativa*

The genomes of the compared *C. sativa* cultivars and landraces ranged in size from 748.82 to 1333.38 Mb (Appendix Table A2), with *C. sativa* DP having the smallest genome and that of Jamaican Lion Dash (JL Dash) being the largest. This variation may be attributed to factors including polyploidy and aneuploidy, which can result in genome duplication or the loss of chromosomal segments (Crawford *et al.,* 2021), differences in the abundance and activity of transposable elements (TEs) and genetic and structural variations between fibre- and drug-type plants. It is more likely however that this variation in genome size is a reflection of the different sequencing and assembly strategies used across the different studies. Assembly artefacts have been shown to inflate and deflate genome size estimates in *Cannabis* (Ryu *et al.,* 2024).

The G+C content of the genomes ranged between 33.49 and 34.24% (Appendix Table A2), with the G+C content of the SG draft genome falling within this range. BUSCO completeness analysis demonstrated that all fifteen compared *C. sativa* genomes fell within the generally accepted range for reference-grade assemblies (above 90%) (Lewin *et al.,* 2022). The genome of *C. sativa* JL Dash cultivar is the most complete genome at 98.2%, with that of *C. sativa* JL being the least complete at only 92.7% (Appendix Table A2).

3.4.1.2. Comparative analysis of repetitive elements in *C. sativa* genomes

Total repeat content detection with RED indicated that approximately 59.24% of the SG genome comprised of repeats (Appendix Table A2). The genomes of *C. sativa* cultivars cs10 and JL Dash incorporated the lowest and highest repeat content, 50.18 and 79.16%, respectively. Analysis of repeats using RepeatMasker trained on the *C. sativa* Pink Pepper genome, indicated that 77.19% of the *C. sativa* SG genome was observed to comprise of repetitive elements (Appendix Table A2). The genome of *C. sativa* cultivar cs10 was once again identified as having the lowest repeat content (64.92%), while the highest repeat content was observed in the genome of *C. sativa* JL Mother (79.16%).

Across all the *C. sativa* genomes, class I retrotransposons were observed to be the most populous repetitive elements, specifically long terminal repeat (LTR) elements, Ty1/Copia elements and Gypsy/DIRS1 elements (Appendix Table A3). Class 1 LTR elements, often referred to as "jumping genes", can move throughout the genome via an RNA intermediate, employing a "copy and paste" mechanism (Havecker *et al.,* 2004). An abundance of class I retrotransposons in a plant genome can lead to an expanded genome size, increased genetic and epigenetic variation, as well as altered gene regulation, which can potentially influence the chemical (cannabinoid profile) and physiological (response to stress) traits of the plant (Havecker *et al.,* 2004; Lanciano & Mirouze, 2017; Wicker *et al.,* 2007). In general, *C. sativa* SG contained fewer class I retrotransposons (430,205 in total, Appendix Table A3) when compared to the other genomes assessed in this study (418911-743012 in total, Appendix Table A3). This suggests that *C. sativa* SG may possesses a more stable and compact genome, a factor that is often sought after by commercial breeders (Hawkins *et al.,* 2009; Vitte *et al.,* 2005). LTR elements contributed 39.97-49.65% (Appendix Table A3) of the repeat content within all the annotated *C. sativa* genomes, with the Finola cultivar containing the highest proportion (49.65%). The *C. sativa* SG draft genome incorporated 47.63% of LTR elements (Appendix Table A3).

Ty1/Copia and Gypsy/DIRS1 elements are both classes of LTR retrotransposons, which are known to be prevalent among plant genomes, contributing to genome size variability (Cheng *et al.,* 2009; Girgis, 2015). Ty1/Copia elements encode proteins such as reverse transcriptases and integrases, which play a role in inducing mutations and structural variations within the genome, and hence they can play key roles in genome evolution (Cheng *et al.,* 2009). Gypsy/DIRS1 elements share structural similarities with Ty1/Copia elements, however, Gypsy/DIRS1 elements differ in their sequence and integration mechanisms which involve

linear, single-stranded intermediates (Malicki *et al.,* 2020). Similarly, Gypsy/DIRS1 elements also contribute to genome diversity and evolution (Piednoël *et al.,* 2011). The Ty1/Copia and Gypsy/DIRS1 elements contributed 23.57-28.53% and 16.36-21.28% (Appendix Table A3) of the repeat content within all the annotated *C. sativa* genomes, respectively. The genome of *C. sativa* Finola cultivar contained the highest proportion of Ty1/Copia elements (28.53%) and that of JL Mother the highest proportion of Gypsy/DIRS1 elements (21.28%). The SG draft genome harboured 27.31% Ty1/Copia elements and 20.27% Gypsy/DIRS1 elements (Appendix Table A3).

Unclassified repeats ranged between 18.60-22.56% (Appendix Table A3). The genome of *C. sativa* JL cultivar contained the highest proportion of unclassified repeats. The SG genome contained a total of 21.81% unclassified repeats (Appendix Table A3), placing it on the upper end of the scale (18.60-22.56 %) when compared with other *Cannabis* genomes in this study.

3.4.1.3. tRNA, rRNA and protein profiles of the *C. sativa* genomes

tRNA profiling showed that between 602 (cs10) and 1,355 (JL Dash) tRNAs are harboured on the *C. sativa* genomes (average: 823.79 tRNAs/genome), with the SG genome containing 648 tRNAs (Appendix Table A2). The most abundant tRNAs in the SG genome were those that carry arginine, leucine and proline. This aligned with the most abundant tRNAs present in *V. vinifera* (the plant gene model used for annotation) and *Arabidopsis*, namely those involved in the transfer of arginine and proline (Petrovic *et al.,* 2019; Yao *et al.,* 2020).

The *C. sativa* genomes harbours between 625 (Abacus) and 3,838 (Cherry Pie) total rRNAs (average: 1146.40 rRNAs/ genome), with the SG genome containing 907 rRNAs (Appendix Table A2). The most rRNAs were predicted within *C. sativa* cultivar Cherry Pie, which has the second smallest genome (Appendix Table A4). Notably, the larger JL genome contains predominantly 5S rRNAs, with only three predicted 18S and six 28S rRNAs (Appendix Table A4).

3.4.1.4. Protein coding sequences in *C. sativa* genomes

On average, the *C. sativa* genomes harbour 51,893 genes, which encode ~50,279 proteins, with the variation reflecting the differences in the genome sizes (Appendix Table A2). The genome of *C. sativa* DP harboured the fewest genes (and protein-coding genes), which corresponds

with its smaller genome (Appendix Table A2), while the genome of *C. sativa* JL Dash harboured 61,156. EggNog-mapper annotated between 34,027 (*C. sativa* DP) and 47,605 (*C. sativa* JL Dash) proteins within the *C. sativa* genomes.

On average 163,795 introns were predicted for the genes identified within the *C. sativa* genomes (and an average of 3.16 introns/ gene) (Appendix Table A2), with the highest number predicted for *C. sativa* JL Dash (208,679) and the genes in *C. sativa* Cherry Pie with the second smallest genome, containing the fewest introns (137,409).

3.4.2. Single-copy orthologue identification and core genome phylogeny construction

A core genome phylogeny of the fifteen compared *C. sativa* genomes was constructed on the basis of 2,085 SCOs (Figure 3.1). The curated alignment comprised 194,629 amino acid positions, of which 1.20% were parsimony-informative (2,341), 3.06% were singletons (5,948) and 95.74% were constant sites (186,340). The core genome phylogeny comprised four clades/clusters, with the genome of *C. sativa* cultivar Finola as outlier in the midpoint rooted tree. *Cannabis sativa* Finola was deliberately bred for the commercial production of hempseed grain and oil, and as such, the cultivar appears is genetically distinct from other *Cannabis* cultivars (Table 3.1) (Finola, 2018; Laverty *et al.,* 2019).

Cluster 1 was comprised of *C. sativa* cultivars Abacus, JL Dash, JL5, JL Mother, JL Father and Purple Kush (Figure 3.1). The cultivars JL Dash, JL5, JL Mother, JL Father all share the same chemotype of Type II (THC:CBD intermediate) with slightly higher CBD levels than THC (Table 3.1) (Appendino *et al.,* 2011). Notably, these four cultivars originate in the USA with plant samples from Marblehead (Massachusetts, USA) (Table 3.1). The *C. sativa* JL5 cultivar (an F1 hybrid cultivar produced by crossing JL Mother and JL Father) clustered closely with both parental genomes. (Figure 3.1). By contrast, the *C. sativa* cultivar Abacus is classified as Type III (CBD dominant fibre-type) and *C. sativa* cultivar Purple Kush is Type I (THC dominant drug-type). The Purple Kush cultivar has genetic origins from *C. sativa* cultivars in the Hindu Kush mountain range between Afghanistan and Pakistan (Clarke & Merlin, 2013). *Cannabis sativa* cultivar Abacus was bred and developed in the USA (with no specific country of origin) (Arsenault *et al.,* 2024).

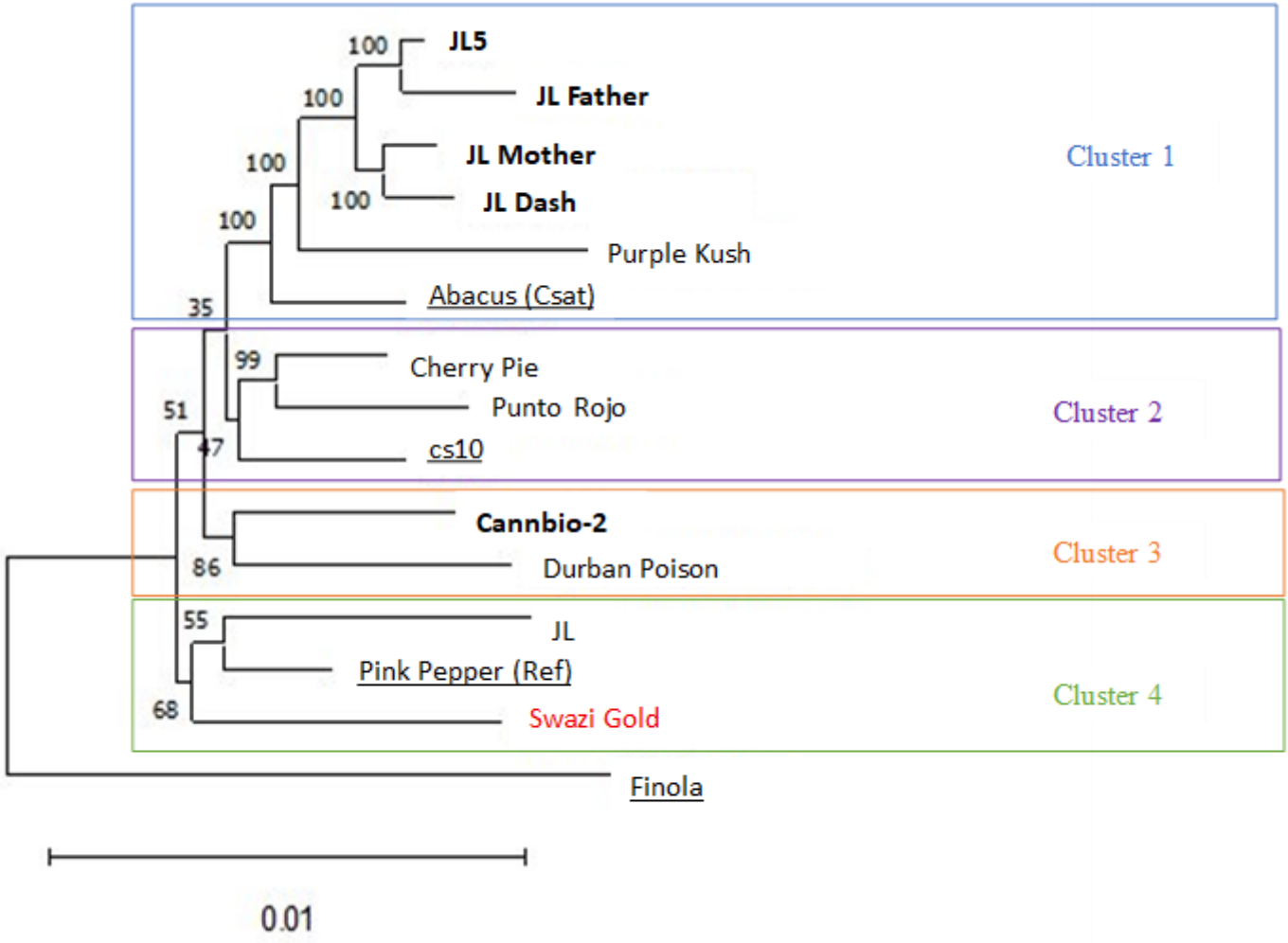


Figure 3.1: Core Maximum likelihood (ML) phylogeny of *Cannabis sativa* cultivars and landraces. This phylogenetic tree indicates the evolutionary relationships between fifteen *C. sativa* genomes including the Swazi Gold genome (marked in red), clustering (marked as Clusters 1, 2, 3 and 4) and chemotype (Chemotype 1 (standard text), **Chemotype 2** (bold text), Chemotype 3 (underlined text)). The core genome phylogeny tree was generated using IQ-Tree3 (Wong *et al.,* 2025) under the best-fit substitution modelVT+F+I+G4, with ultrafast bootstrapping (1000 replicates) (Hoang *et al.,* 2018).

Cluster 2 consisted of *C. sativa* cultivar Cherry Pie (Type I), landrace Punto Rojo (Type I) and cultivar cs10 (Type III) (Figure 3.1) (Table 3.1). *Cannabis sativa* Cherry Pie (a so called *indica*-dominant hybrid) was produced by cross breeding *C. sativa* Grandaddy Purple and Durban Poison, with the purpose of developing a potent plant with high THC levels and low CBD levels (https://www.cannaconnection.com/strains/cherry-pie). *Cannabis sativa* Punto Rojo is a South American landrace originating from the coastal and central highland regions of Colombia (https://app.jointcommerce.com/blog/category/cannabis-101/punto-rojo-by-the-

landrace-team-a-comprehensive-strain-guide/). *Cannabis sativa* cs10 is a hybrid cultivar, a cross between the Skunk#1 (drug-type) and Carmen (fibre-type) cultivars, developed for a high CBD:THC ratio (Laverty *et al.,* 2019).

The smaller cluster 3 contained *C. sativa* landrace DP and cultivar Cannbio-2, which are both classified as chemotype I (THC dominant drug-type) (Table 3.1) (Braich *et al.,* 2020). The *C. sativa* DP landrace originates from South Africa (Durban, KwaZulu-Natal) and is regarded as a pure *sativa* drug-type plant cultivated for its high THC levels (Table 3.1) (DNA Genetics, 2025). *Cannabis sativa* cultivar Cannbio-2 was developed in Australia to produce a balanced THC:CBD ratio (Grassa *et al.,* 2021).

Cannabis sativa landrace SG was located within the fourth cluster alongside *C. sativa* landrace JL and cultivar Pink Pepper (Figure 3.1.). *Cannabis sativa* Pink Pepper is classified as a chemotype III (CBD dominant fibre-type) plant (Table 3.1) (Tomas & ElSohly, 2015), while *C. sativa* SG is classified as chemotype I (THC dominant drug-type) (Crampton, 2015). The *C. sativa* JL cultivar has unknown CBD and THC levels, and therefore has an unknown chemotype (Gao *et al.,* 2020). Both *C. sativa* JL (collected in China - Jilong, Tibet) and SG (collected in South Africa) are wild-type plants cultivated and selected over several plant generations with limited hybridisation from exogenous *C. sativa* strains (Crampton, 2015; Gao *et al.,* 2020). By contrast, *C. sativa* cultivar Pink Pepper was bred from 26 cultivars during a collaboration between universities in South Korea and the USA (https://www.instituteofcannabisresearchcolorado.org/pink-pepper/) but the exact cultivars that were cross-bred to produce Pink Pepper are not stated in literature (Ryu *et al.,* 2024) Therefore, we cannot exclude the possibility that both *C. sativa* JL and *C. sativa* SG may have been included within these 26 cultivars.

The core genome phylogeny showed clustering of plants on the basis of their chemotype for only one clade, namely Cluster 1. The other clusters comprised of *C. sativa* cultivars and landraces with distinct chemotypes. This may be due to the polygenic nature of *C. sativa* chemotypes, epigenetic regulation, mobile genetic elements, copy number variation (CNV), hybridization/selective breeding or genomic introgression without genomic restructuring (Laverty *et al.,* 2019; Lynch *et al.,* 2016; McKernan *et al.,* 2020; Pacifico *et al.,* 2006; Vergara *et al.,* 2019).

The chemotype of a *C. sativa* plant is influenced by multiple genes (polygenic), which are controlled through epigenetic regulatory mechanisms (DNA methylation, histone

modification), ultimately impacting cannabinoid gene expression (Lynch *et al.,* 2016; McKernan *et al.,* 2020). Subsequently, key biosynthetic genes such as *THCAS* and *CBDAS* may be located on mobile genetic elements and exhibit CNV, influencing plant chemotype independently of core genome structure (Innes & Vergara, 2022; McKernan *et al.,* 2020). Additionally, extensive hybridization and selective breeding between genetically distant lines (fibre- vs drug-type) has introduced desired chemotype traits across lineages (as indicated by Pink Pepper produced using 26 cultivars) (Laverty *et al.,* 2019). This ultimately leads to introgressed chemotype traits from one lineage to another, without affecting its overall genome structure (Pacifico *et al.,* 2006).

ML phylogeny based on SCOs indicates core genomic relationships but overlooks structural and regulatory gene variations such as the occurrence of CNVs among these elements. On this basis, cannabinoid profiles cannot be reliably linked to evolutionary history and that rather extensive laboratory cross-breeding and selective breeding for the cannabinoid profiles lies at the basis of the chemotypes.

3.4.3. Analysis of cannabinoid biosynthetic genes in *C. sativa*

A tBlastn analyses with reference proteins involved in cannabinoid biosynthesis revealed the presence of 10/11 genes encoding enzymes that comprise the cannabinoid biosynthetic pathway within all fifteen annotated genomes (Innes & Vergara, 2023) (Appendix Table A5). The presence of these genes within all the annotated genomes, including the genome of *C. sativa* SG, indicated the potential of these plants to produce cannabinoids (Innes & Vergara, 2023). Aromatic prenyltransferase (AP/CBGAS) was not detected within any of the *C. sativa* genomes (Appendix Table A5). This may be attributed to challenges encountered during local BLAST analysis for *AP* gene identification, primarily due to the absence of a well-characterised and validated reference protein sequence in databases including the NCBI (www.ncbi.nlm.nih.gov/) and Uniprot (www.uniprot.org/) databases. Therefore, the presence of *AP* within the genomes cannot be ruled out, and further research would need to be conducted to conclusively confirm its absence.

NJ phylogeny of the aligned DXS protein sequences shows the cannabinoid biosynthesis begins with a polyketide synthetic pathway, where acetyl-CoA carboxylase (ACCase) plays a role in the conversion of acetyl-CoA to malonyl-CoA (Dewick, 2009; Gagne *et al.,* 2012). A tBlastn analysis identified 35 homologues encoded on the comparator genomes, averaging two

paralogous protein copies/genome, with five copies in *C. sativa* Purple Kush. The ACCase proteins form two clusters in the NJ tree (Figure 3.2). The *C. sativa* SG genome contains two distinct ACCase paralogues, which shared 85.9% amino acid identity and formed part of the two distinct clades, with *C. sativa* SG ACCase copy 1 sharing 96.2% amino acid identity with the ACCase reference protein (NCBI XM_030629499), and the second copy sharing 87.6% amino acid identity.

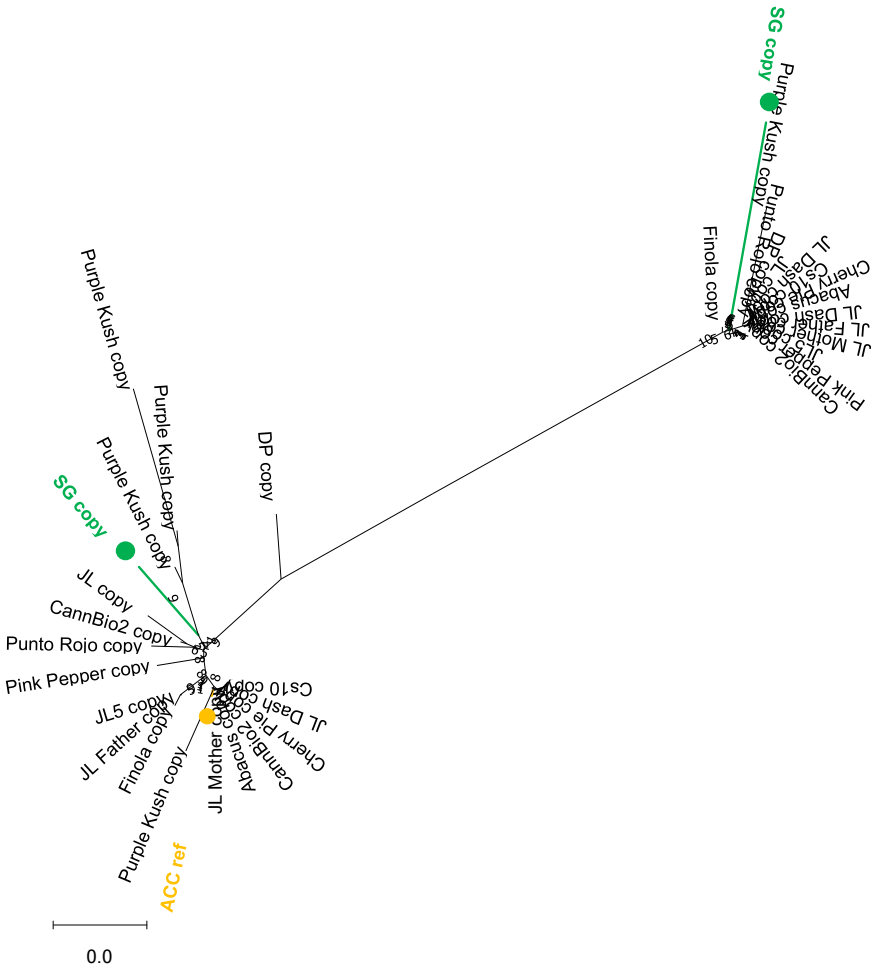


Figure 3.2: Neighbour-Joining (NJ) phylogenetic tree of acetyl-CoA carboxylase (ACCase) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the ACCase protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

Alongside the conversion of acetyl-CoA to malonyl-CoA, hexanoyl-CoA synthetase (HCAS) catalyses the conversion of hexanoic acid to hexanoyl-CoA thereby, initiating cannabinoid

biosynthesis (Gagne *et al.,* 2012). A total of 67 HCAS homologues were present in the comparator genomes, with an average of five paralogues per genome. Most paralogous copies (six) were observed in three cultivars, *C. sativa* CannBio2, JL Father and JL Dash, while three copies each were observed in *C. sativa* Cherry Pie and Finola. The HCAS paralogues occur in four distinct clusters in the NJ tree (Figure 3.3). There are four distinct paralogues in *C. sativa* SG, sharing an average amino acid sequence identity of 56.6% among them (range 44.2-77.8%). These paralogous copies fall within each of the four clades, with *C. sativa* SG HCAS copy 1 sharing 99.3% amino acid identity with the HCAS reference protein (GenBank MN551226), while SG HCAS copy 2 shares 77.9% amino acid identity.

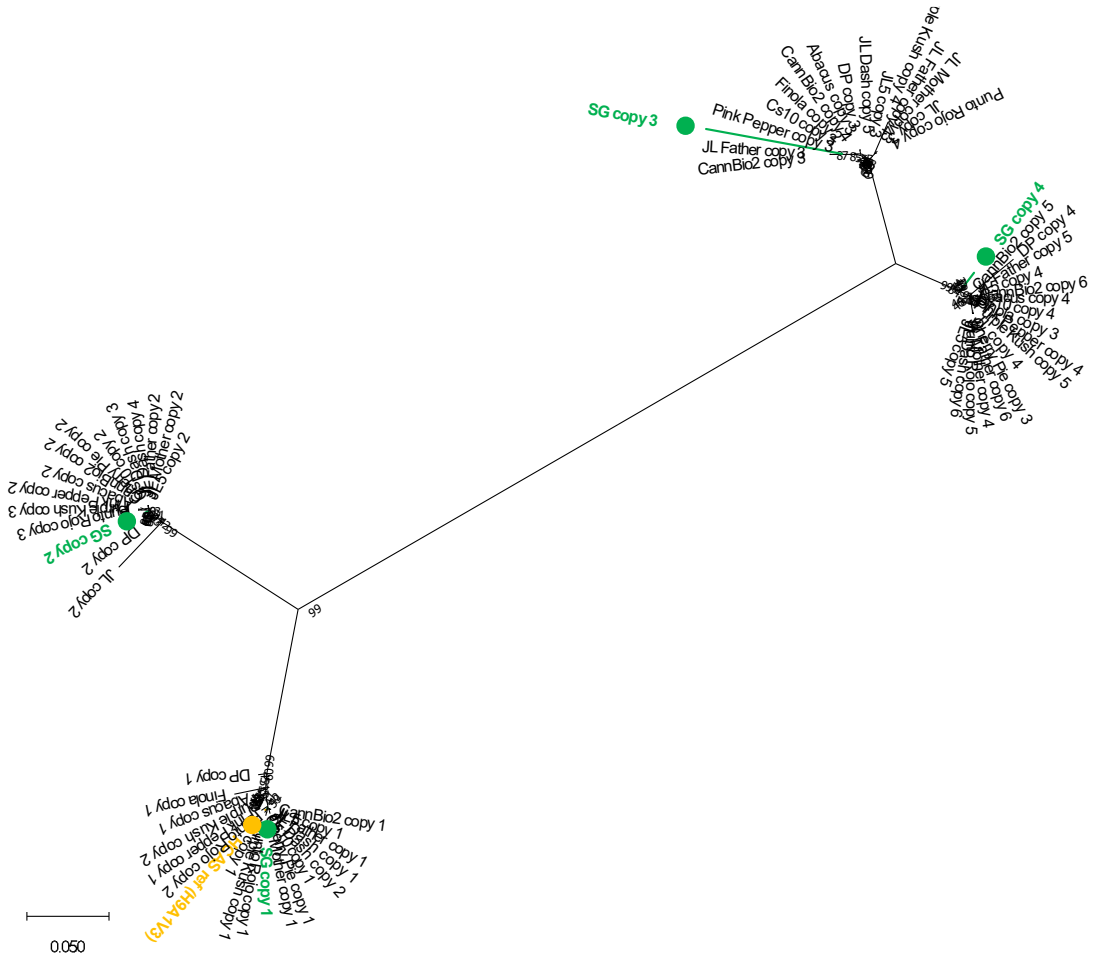


Figure 3.3: Neighbour-Joining (NJ) phylogenetic tree of hexanoyl-CoA synthetase (HCAS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the HCAS protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

In the polyketide synthetic pathway olivetol synthase / tetraketide synthase (OLS/TKS) catalyses the condensation of hexanoyl-CoA (to produce olivetolic acid - OLA) or butanoyl-CoA (to produce divarinic acid) with malonyl-CoA, to produce a linear tetraketide intermediate (Dewick, 2009; Gagne *et al.,* 2012). A tBlastn analysis identified a total of 80 homologues among the comparator genomes, with five paralogous copies on average encoded per genome. High variability in copy number was observed, with only three paralogues in *C. sativa* landrace DP and nine copies in JL Father, while six paralogous copies occur in *C. sativa* landrace SG. An NJ phylogeny on the basis of the aligned OLS/TKS protein sequences (Figure 3.4.) demonstrates the OLS/TKS paralogues form six distinct clusters, with the *C. sativa* SG olivetol OLS/TKS copies occupying five of these clusters. These *C. sativa* SG OLS/TKS protein copies shared an average amino acid identity of 69.3% (range 64.2-88.7%). Overall, two OLS/TKS copies of *C. sativa* SG cluster with the reference OLS/TKS protein (GenBank AB164375.1), sharing 97.1% and 89.7% amino acid sequence identity with the reference protein for *C. sativa* OLS/TKS copy 1 and copy 2, respectively.



Figure 3.4: Neighbour-Joining (NJ) phylogenetic tree of tetraketide synthase (OLS/TKS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the OLS/TKS protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

The linear tetraketide intermediate undergoes condensation to form OLA, a process induced by the enzyme olivetolic acid cyclase (OAC) (Gagne *et al.,* 2012; Taura *et al.,* 2007). The comparator genomes contained 26 homologues for OAC, with only a single copy in five of the compared, two paralogous copies in seven and three copies in three of the compared *C. sativa* cultivars (Figure 3.5). The OAC homologues are highly conserved, sharing 98.3% average amino acid identity between them across the 101 amino acid reading frame.

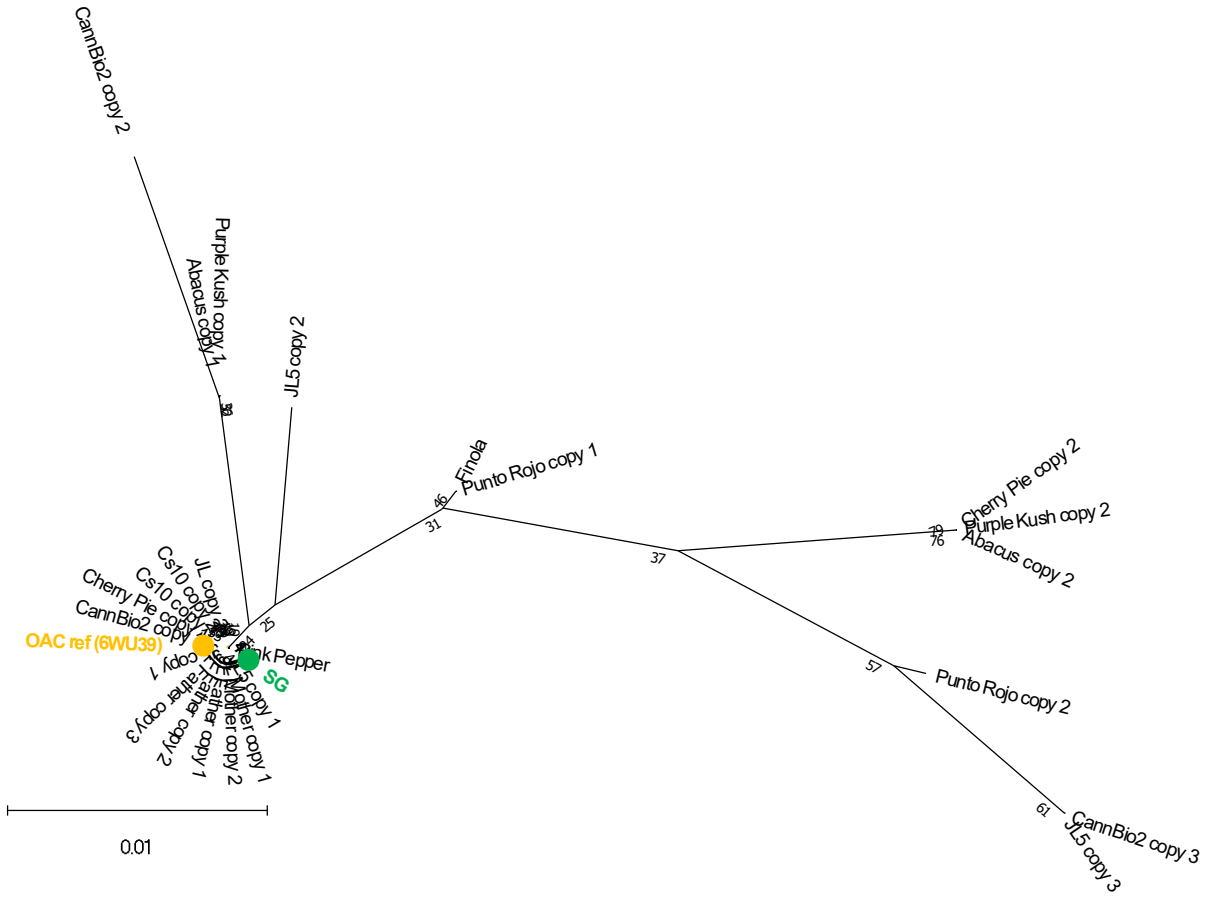


Figure 3.5: Neighbour-Joining (NJ) phylogenetic tree of olivetolic acid cyclase (OAC), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the OAC protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

The second pathway in cannabinoid biosynthesis, the methylerythritol-4-phosphate (MEP) pathway, is initiated by 1-deoxy-D-xylulose-5-phosphate synthase (DXS) to produce isopentenyl diphosphate and dimethyl allyl diphosphate (Xie *et al.,* 2023). There are two isoforms of this enzyme, DXS1 and DXS2 in *C. sativa* (Tian *at el*., 2022), sharing 57.9% amino acid identity. A total of 57 homologues were retrieved from the genomes of the comparator *C. sativa* cultivars, with four paralogous copies/genome. NJ phylogeny of the aligned DXS protein sequences shows the paralogues clustering in three clades (Figure 3.6.), with the two reference isoforms forming part of two of these clades. *Cannabis sativa* SG DXS copy 1 shares 100% amino acid identity with the reference DXS1 protein (NCBI XM_061104475.1), while SG copy 2 shares 99.5% amino acid identity with DXS2 (GenBank KY014577.1) The third

C. sativa SG DXS paralogue only shares 52.5% amino acid identity with SG DXS copy 1 and 2.

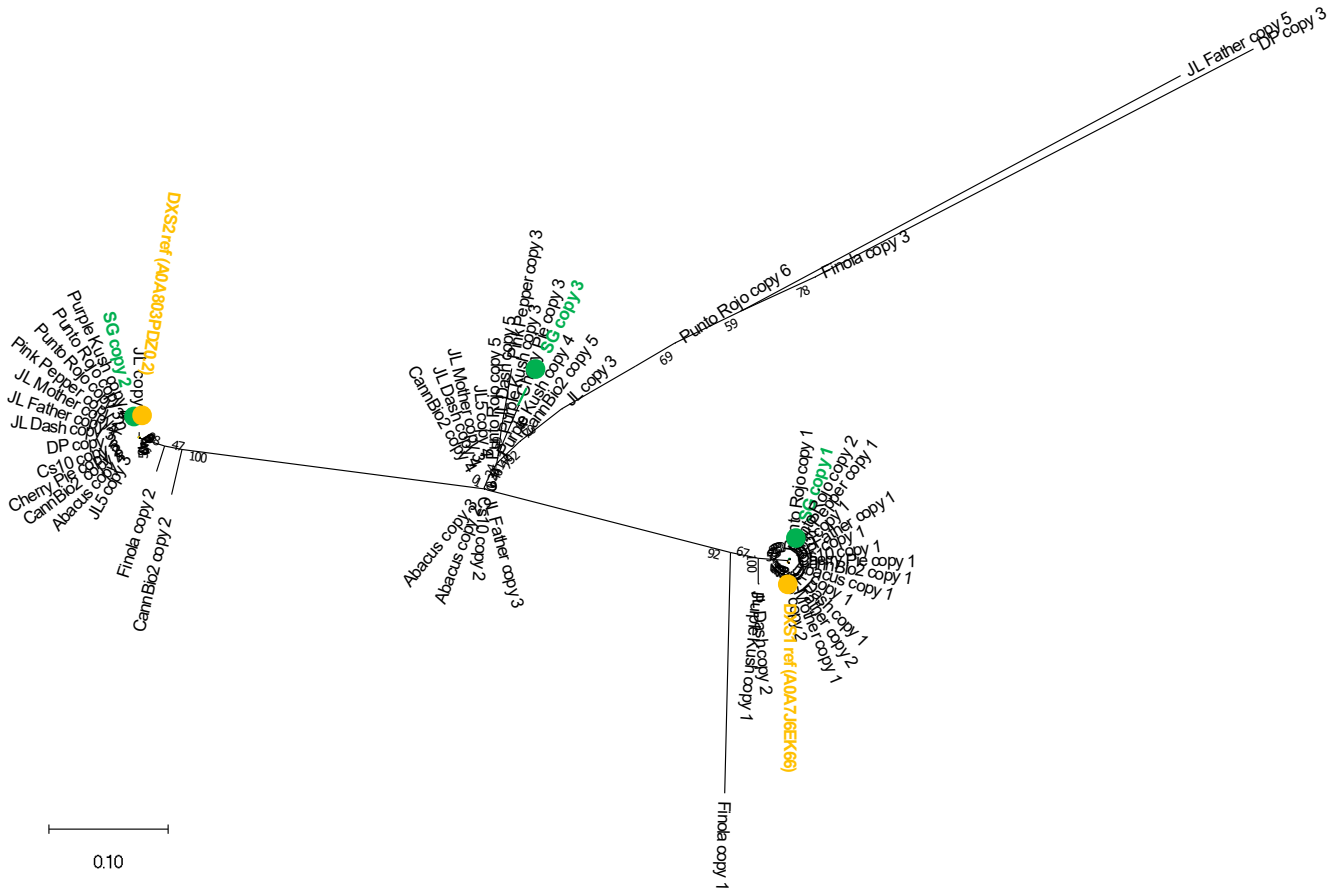


Figure 3.6: Neighbour-Joining (NJ) phylogenetic tree of 1-deoxy-D-xylulose-5-phosphate synthase (DXS) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the DXS protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

Within the MEP pathway geranyl pyrophosphate synthase (GPPS) produces geranyl pyrophosphate (GPP), an important monoterpene fraction needed for cannabinoid biosynthesis (Gulck & Moller, 2020). Similar to 1-deoxy-D-xylulose-5-phosphate synthase, there are two isoforms (GPPS1 and GPPS2), sharing 55.3% amino acid identity and the homologues of both isoforms form two distinct clusters in the resultant NJ phylogeny (Figure 3.7). Most *C. sativa* genomes contain one copy each of the genes coding for GPPS1 and GPPS2, except for *C. sativa* JL and CannBio2 which have two copies of isoform GPPS1 and JL Dash and JL Father, which contain two copies of each isoform. The additional gene copies in these cultivars may reflect

their larger genome sizes (Crawford *et al.,* 2021) (Appendix Table A2), however further analyses would be required. The *C. sativa* SG isoforms share 100 and 97.1% sequence identity to GPPS1 (GenBank KY014573.1) and GPPS2 (GenBank KY014574.1), respectively.

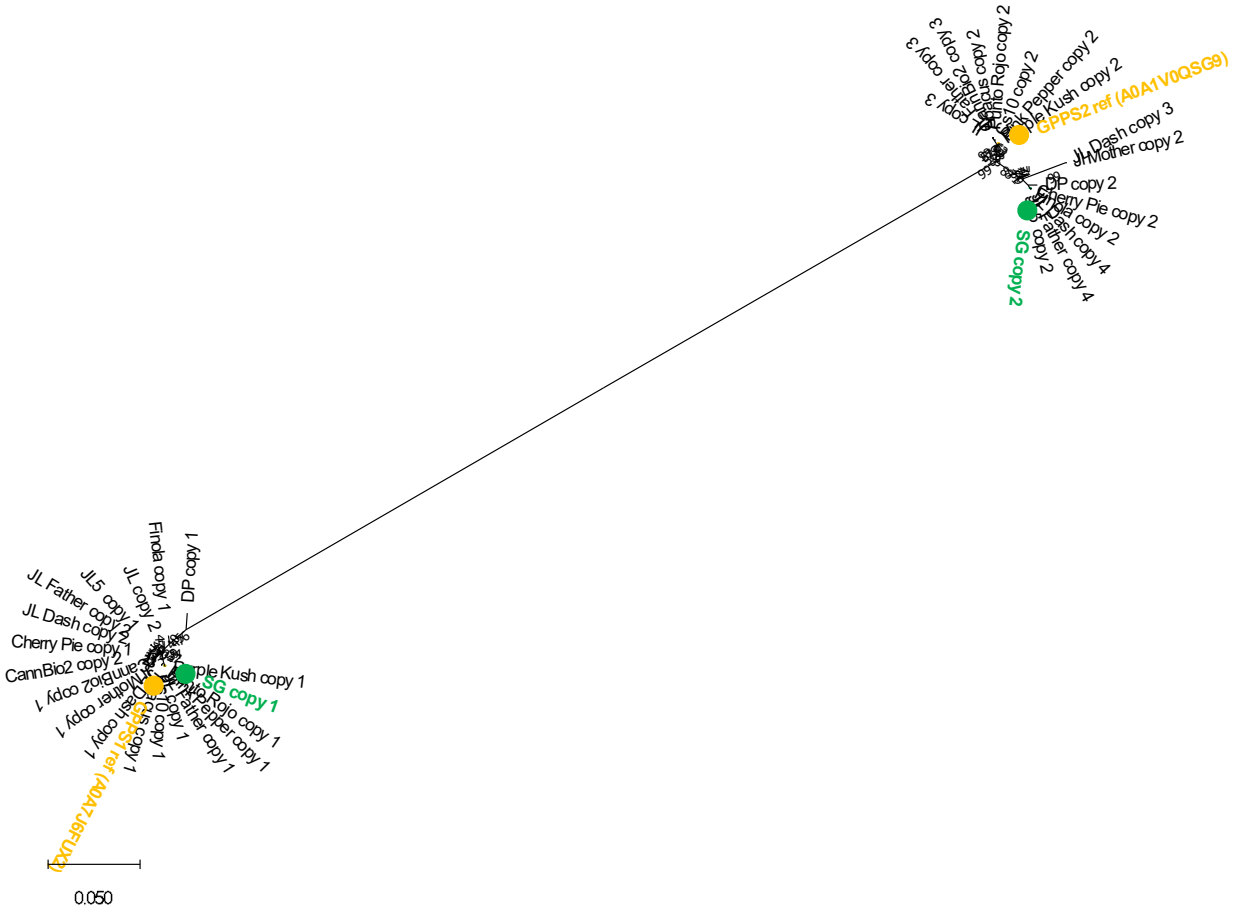


Figure 3.7: Neighbour-Joining (NJ) phylogenetic tree of geranyl pyrophosphate synthase (GPPS), protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the GPPS protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018).

The geranylpyrophosphate olivetolate geranyltransferase (GOT) enzyme transfers a geranyl group from GPP to OLA to form cannabigerolic acid (CBGA), the precursor from which all cannabinoids are produced (Luo *et al.,* 2019). An NJ tree based on the 42 aligned GOT homologues comprised five clusters, indicating various divergent events for this gene among the genomes (Figure 3.8.). An average of three paralogues per comparator genome were identified with two paralogues in *C. sativa* SG and four paralogous copies each in *C. sativa* JL Dash, JL Father, Purple Kush and Cherry Pie. The two *C. sativa* SG paralogues shared only

21.8% sequence identity and only 34.2 and 19.6% amino acid identity with the reference GPPS protein (NCBI XM_061106831.1), supporting extensive sequence diversification for this enzyme.

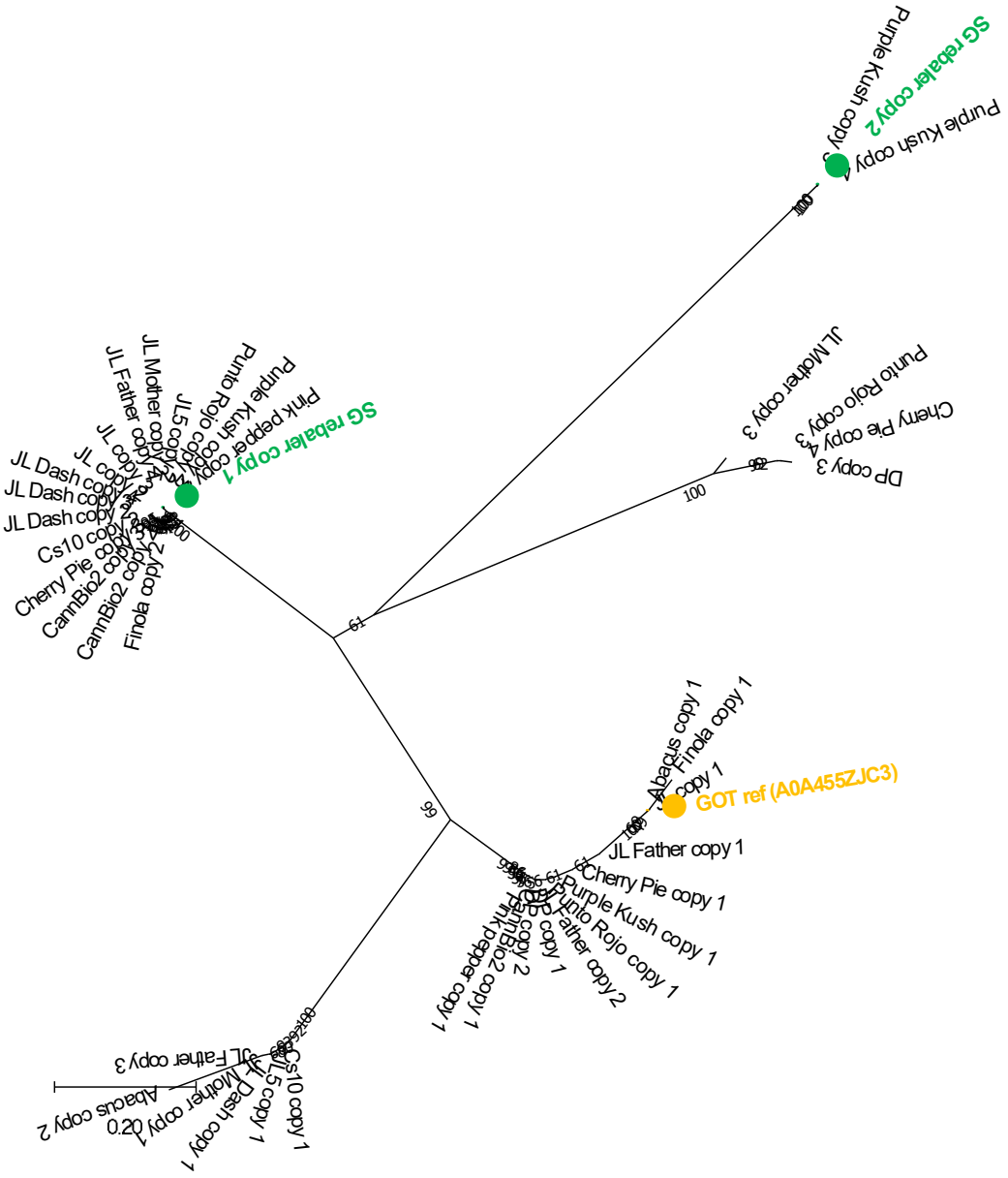


Figure 3.8: Neighbour-Joining (NJ) phylogenetic tree of geranylpyrophosphate olivetolate geranyltransferase (GOT) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the GOT protein reference sequence in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.*, 2018).

During cannabinoid biosynthesis tetrahydrocannabinolic acid synthase (THCAS), cannabidiolic acid synthase (CBDAS) and cannabichromenic acid synthase (CBCAS) are crucial for the synthesis of cannabichromenic acid (CBCA), cannabidiolic acid (CBDA) and Δ^9-tetrahydrocannabinolic acid (Δ^9-THCA) (the carboxylic acid forms that cannabinoids are stored in) from cannabichromenic acid (CBCA), respectively (Marks *et al.,* 2009; Taura *et al.,* 1996). Notably, single nucleotide polymorphisms (SNPs) differentiating between *THCA* and *CBDA* synthase genes have been linked to separate chemotypes I, II and III (Pacifico *et al.,* 2006). The reference proteins for THCAS, CBCAS and CBDAS (JQ437488.1; MW382907.1 and XM_061118802.1, respectively) share 87-93% amino acid identity (Laverty *et al.,* 2019), hence analyses of these proteins were combined. A total of 107 homologues were found to be encoded on the comparator *C. sativa* genomes, with an average number of seven paralogous copies/genome. The lowest number of paralogues were identified in *C. sativa* landrace DP (three copies), while the most copies were observed in *C. sativa* JL Father (11 copies). This suggests that the genes coding for these proteins are subject to extensive gene duplication. Alignment of the amino acid sequences and a resultant NJ phylogeny showed that the THCAS, CBCAS and CBAS orthologues form eight distinct clades, with the reference proteins of all three enzymes forming part of a single cluster (Figure 3.9). Five distinct paralogues were observed in *C. sativa* SG. These formed part of four distinct clades in the phylogeny. The *C. sativa* SG paralogues share 52.92% (range 45.1-63.7%) amino acid identity among the copies, suggesting far greater divergence in these central cannabinoid enzymes than previously observed (van Bakel *et al.,* 2011). They further displayed 50.5% average amino acid identity with the reference THCAS, CBCAS and CBDAS proteins.

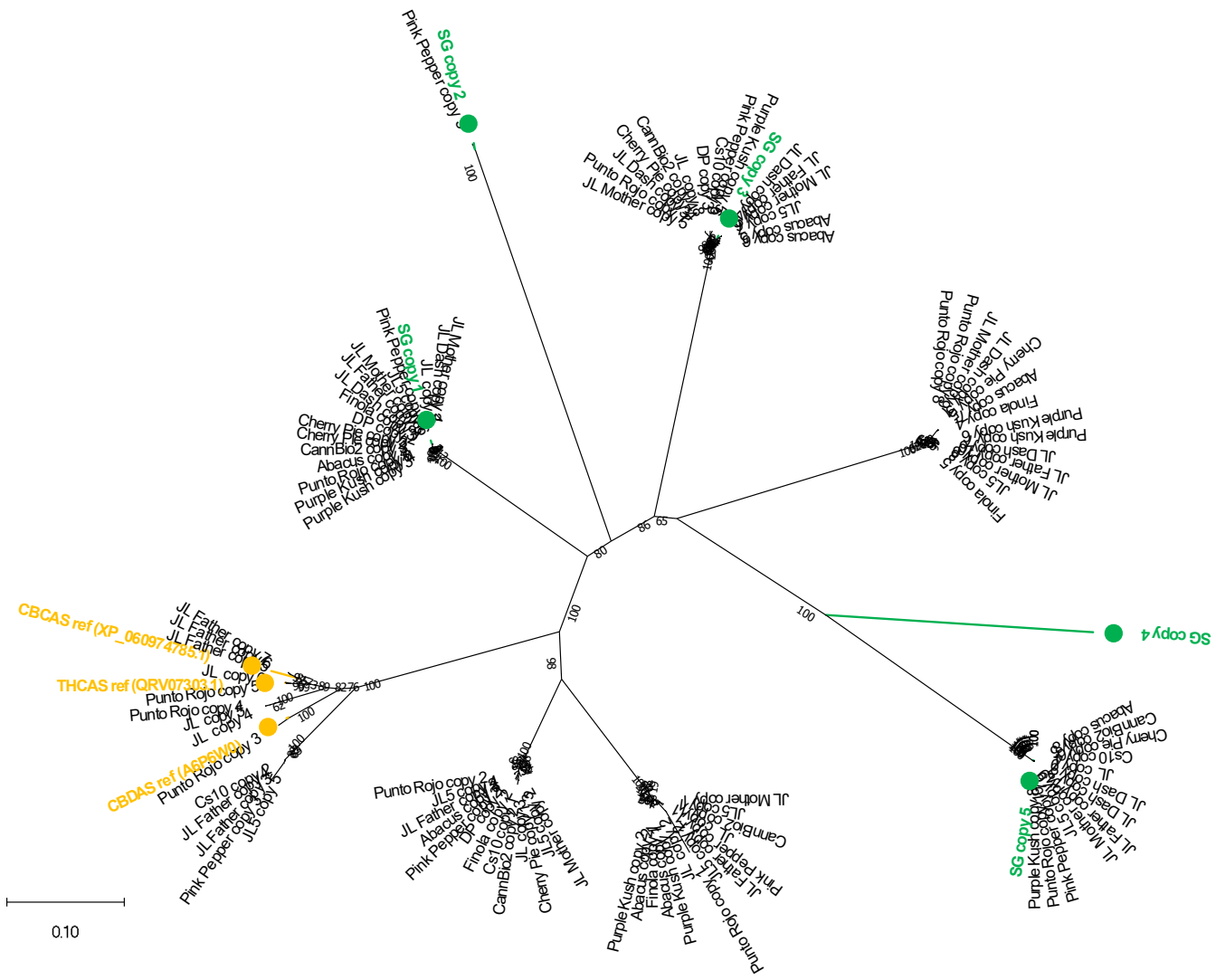


Figure 3. 9: Neighbour-Joining (NJ) phylogenetic tree of cannabidiolic acid synthase (CBDAS), cannabichromenic acid synthase (CBCAS) and tetrahydrocannabinolic acid synthase (THCAS) protein paralogues identified across fifteen annotated *Cannabis sativa* genomes. *C. sativa* Swazi Gold (SG) paralogues are highlighted in green and the CBDAS, CDCAS and THCAS protein reference sequences in orange. The tree was constructed using the test NJ method in MEGA12 (Kumar *et al.,* 2018)

Overall, the analyses of the individual cannabinoid biosynthetic proteins revealed differentiated evolutionary trajectories for each enzyme. Varying degrees of paralogy were evident for all the enzymes, except in the case of DXS and GGPS that are known to occur as two isoforms (Table 3.2). This is exemplified by the THCAS, CBCAS and CBDAS enzymes in *C. sativa* JL Father, where eleven paralogous copies exist in its genome. This makes it difficult to assume which of the copies play a role in the production of these different

cannabinoids, namely Δ^9-tetrahydrocannabinol (Δ^9-THC), and the non-psychotropics cannabidiol (CBD), cannabichromene (CBC), and cannabigerol (CBG), or whether they are in fact involved in the synthesis of hereto unknown and novel phytocannabinoids. Further research would be needed to establish whether the paralogy resulted in the sub-functionalisation (division of labour) or the neo-functionalisation (novel function) of each enzyme copy.

The high levels of paralogy further complicates the delineation of *C. sativa* cultivars as drug- or fibre-type chemotypes or the chemotypes that embrace both sides. It is evident that this paralogy may have been introduced through introgression of mobile cannabinoid biosynthetic genes via extensive crossbreeding between fibre- and drug-type plants. Such introgression may have led to increased numbers of copies of *THCAS*, shifting cannabinoid biosynthesis toward THC production (leading to drug type phenotype) without substantially altering the overall genomic structure, as indicated by shared elements across chemotypes I, II and III (Pacifico *et al.,* 2006). Consequently, chemotype determination in *C. sativa* is much more complex than previously expected, being influenced by multiple genomic and evolutionary factors, including gene introgression through crossbreeding and hybridization, the polygenic nature of cannabinoid biosynthetic genes, the possible location of these genes on mobile genetic elements and CNV (Laverty *et al.,* 2019; Lynch *et al.,* 2016; McKernan *et al.,* 2020; Pacifico *et al.,* 2006; Vergara *et al.,* 2019). Further research should investigate the functional roles of these cannabinoid enzymes and their CNV, to dissect how this genomic architecture contributes to the ultimate cannabinoid profile and phenotype of the plant.

3.4.4. Biosynthetic gene cluster analysis

Biosynthetic gene clusters (BGCs) were identified in all fifteen *C. sativa* genomes using PlantiSMASH (Kautsar *et al.,* 2017). A total of 501 BGCs were predicted, with an average of 35.79 per genome (Appendix Table A6). The *C. sativa* Finola and Purple Kush genomes contained the most BGCs (42) and that of Punto Rojo contained the fewest (27). Across the comparator *C. sativa* genomes saccharide BGCs (range 5-12) were the most abundant, followed by alkaloids (range 4-7) and terpenes (range 3-6); and the least common BCGs were lignan (2) and cyclopeptide BGCs (3) (Appendix Table A6). Analysis of the *C. sativa* SG draft genome identified 31 BCGs, predominantly those for saccharide metabolites. These BGC

abundance patterns are consistent with reported metabolic diversity within *C. sativa* (Zager *et al.,* 2019).

Although they are the most abundant BGCs within *C. sativa,* saccharide secondary metabolites are not well understood or characterised (ElSohly & Gul, 2014). Saccharide BGC refers to the clusters of genes encoding the enzymes and components needed to synthesize specialized carbohydrate-derived molecules (sugar- derived specialized secondary metabolites) (Polturak & Osbourn, 2021). However, saccharide metabolism is often part of the plants primary metabolic pathways involving genes spread throughout the genome rather than being focused into gene clusters (Louveau & Osbourn, 2019). The identified saccharide BGCs play roles in the synthesis of glycosyltransferases, regulators, transporters and modifying enzymes (Polturak & Osbourn, 2021; Smit *et al.,* 2022). These clusters have been theorised to contribute to the transportation of bioactive compounds and the tolerance of plants to stress and environmental conditions (Louveau & Osbourn, 2019). Further research into these BGCs in *C. sativa* and their impact on the biosynthetic pathways is required.

The highly valued aromas and flavours of *C. sativa* are attributed to the production of terpenes (Booth & Bohlmann, 2019). Well-known terpenes in *C. sativa* include myrcene (earthy notes), limonene (citrus notes), pinene (pine notes), linalool (floral notes) and β-caryophyllene (spicy notes) (Booth & Bohlmann, 2019; Russo, 2011). Terpinolene typically contributes to fruity and floral aromas which have been associated with improving the sensory qualities of *C. sativa*. It is a monoterpene with pharmacological characteristics including antioxidant and mild sedation (de Sousa *et al.,* 2007; Russo, 2011). In contrast, myrcene is associated with sweeter more earthy aromas (Russo, 2011) and affects the human central nervous system (CNS), exerting mild sedation and muscle relaxation (Jansen *et al.,* 2019). Pinene provides a sharp, herbal, pine-like aroma and has been reported to enhance alertness and cognitive abilities (Miyazawa *et al.,* 2011; Booth & Bohlmann, 2019; Russo, 2011). Within the *C. sativa* SG genome, terpene BCGs associated with terpinolene, myrcene and pinene biosynthesis were identified (Appendix Table A6). This terpene profile aligns with the herbal aroma and sweet, earthy flavour associated with this landrace (The Landrace Team; https://www.tltseeds.com/).

Alkaloid BGCs refer to the gene clusters that are involved in the synthesis of specialized nitrogen-containing metabolites (Smit *et al.,* 2022). Alkaloid BGCs in *C. sativa* are also not well understood or characterised; those that have been identified are associated with plant defence (cannabisativine and cannabimine C) and growth regulation (ElSohly & Gul, 2014, Hourfane *et al.,* 2023). Within the *C. sativa* SG genome, alkaloid BGCs were identified

containing key biosynthetic domains such as Bet_v_1, Cu_amine_oxid, Str_synth, BBE, Orn_DAP_Arg_deC, and Pyridoxal_deC (Kautsar *et al.,* 2017). These domains are involved in a variety of enzymatic functions including alkaloid synthesis, oxidative reactions and structural modification (Kautsar *et al.,* 2017; Rea *et al.,* 2004; Winkler *et al.,* 2008). The presence of these alkaloid BGCs within *C. sativa* SG suggests the potential production of alkaloid-like compounds (such as nitrogenous metabolites) involved in plant defence or environmental adaptation (Kautsar *et al.,* 2017; Rea *et al.,* 2004).

3.5. Conclusion

This study conducted a comparison of *C. sativa* genomes with varying chemical profiles. Core genome phylogenetic analysis revealed unexpected clustering, not in alignment with either the chemotype profile or the geography of the compared cultivars. BLAST-analysis confirmed the presence of cannabinoid biosynthesis genes across all fourteen comparator *C. sativa* genomes. Comparison of the cannabinoid biosynthetic enzymes, however, revealed evidence of extensive paralogy for most of the enzymes, regardless again of phenotype or geographic origin. These factors can likely be attributed to decades of clandestine and unregulated crossbreeding of this controversial plant, resulting in expansive gene introgression, and concomitantly, altered chemotypes, with the potential production of a large set of cannabinoids which have thus far not been identified using chemical means. The BCGs identified in this study provide a foundation for future comparative genome analyses into gene expression patterns of BGCs and cannabinoid biosynthetic genes within *C. sativa* across chemotypes. Additional exploration into the divergent paralogues and CNVs within *C. sativa* genomes may reveal novel or hybrid clusters worth exploring for possible industrial or therapeutic application. Overall, these findings underscore the polygenic and dynamic nature of cannabinoid biosynthesis, making chemotype determination substantially more complex than initially expected. Furthermore, the data highlights the need for research focused on exploring the regulatory mechanisms and gene CNVs that shape chemotype expression.

3.6. References

Andre, C.M., Hausman, J.F., Geurriero, G. (2016). '*Cannabis sativa*: The plant of the thousand and one molecules', *Frontiers in Plant Science*, 7, pp. 1-14. Doi: 10.3389/fpls.2016.00019

Appendino, G., Chianese, G., Taglialatela-Scafati, O. (2011). 'Cannabinoids: Occurrence and medicinal chemistry', *Current Medicinal Chemistry*, 18 (7), pp. 1085-1099. Doi: 10.2174/092986711794940888

Arsenault, T.L., Prapayotin-Riveros, K., Ammirata, M.A., White, J.C., Dimkpa, C.O. (2024). 'Compliance Testing of Hemp (*Cannabis sativa* L.) Cultivars for total delta-9 THC and total CBD using gas chromatography with flame ionization detection', *Plants*, 13(4), pp.519. Available at: https://doi.org/10.3390/plants13040519 (Accessed 30/09/2025).

Aviello, G., Romano, B., Borrelli, F., Capasso, R., Gallo, L., Piscitelli, F., Marzo, V.D., Izzo, A.A. (2012). 'Chemopreventive effect of the non-psychotropic phytocannabinoid cannabidiol on experimental colon cancer', *Journal of Molecular Med*icine, 90, pp.925-934. Doi: 10.1007/s00109-011-0856-x

Bairoch, A., Apweiler, R. (2000). 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Research*, 28(1), pp.45-48. Doi: 10.1093/nar/28.1.45

Barrick, J.E., Colburn, G., Deatherage, D.E., Traverse, C.C., Strand, M.D., Borges, J.J., Knoester, D.B., Reba, A., Meyer, A.G. (2014). 'Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq', *BMC Genomics*, 15, pp.1039. Doi: 10.1186/1471-2164-15-1039s

Bagheri, M., Mansouri, H. (2015). 'Effect of induced polyploidy on some biochemical parameters in *Cannabis sativa* L.', *Applied Biochemistry and Biotechnology*, 175, pp. 2366-2375. Doi: 10.1007/s12010-014-1435-8.

Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., Mayer, K.F.X. (2014). 'Plant genome sequencing — applications for crop improvement', *Current Opinion in Biotechnology*, 26, pp.31–37. Doi: 10.1016/j.copbio.2013.08.019

Booth, J.K., Bohlmann, J. (2019) 'Terpenes in *Cannabis sativa* – From plant genome to humans', *Plant Science*, 284, pp. 67-72. Doi: 10.1016/j.plantsci.2019.03.022.

Bradshaw, R.H.W., Coxon, P., Greig, J.R.A., Hall. A.R. (1981). 'New fossil evidence for the past cultivation and processing of hemp (*Cannabis sativa* L.) in Eastern England', *New Phytologist*, 89, pp. 503-510. Doi: 10.1111/j.1469-8137.1981.tb02331.x

Braich, S., Baillie, R.C., Spangenberg, G.C., Cogan, N.O.I. (2020). 'A new and improved genome sequence of *Cannabis sativa'*, *Gigabyte*, pp. 1-13. Doi: 10.46471/gigabyte.10s

Carley, D.W., Prasad, B., Reid, K.J., Malkani, R., Attarian, H., Abbott, S.M., Vern, B., Xie, H., Yuan, C., Zee, P.C. (2018). 'Pharmacotherapy of apnea by cannabimimetic enhancement, the PACE clinical trial: Effects of dronabinol in obstructive sleep apnea', *Sleep*, 41(1). Doi: 10.1093/sleep/zsx184

Castresana, J. (2000). 'Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis', *Molecular Biology and Evolution*, 17 (4), pp. 540-552. Doi: 10.1093/oxfordjournals.molbev.a026334

Chaney, L., Sharp, A.R., Evans, C.R., Udall, J.A. (2016). 'Genome mapping in plant comparative genomics', *Trends in Plant Science*, 21(9), pp.770–780. Doi: 10.1016/j.tplants.2016.05.004

Cheng, X., Wang, J., Zhang, X., Zhang, X. (2009). 'A new family of Ty1-copia-like retrotransposons identified in the genome of *Gossypium hirsutum*', *BMC Genomics*, 10(1), p. 1. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC2666490/ (Accessed 17/10/2025)

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., Korlach, J. (2013). 'Nonhybrid, finished microbial genome assemblies using long-read sequencing technology', *Nature Methods*, 10(6), pp.563-569. Doi: 10.1038/nmeth.2474

Clarke, R.C. (1981). 'The genetics and breeding of *Cannabis*', in *Marijuana Botany: An Advanced Study: The Propagation and Breeding of Distinctive Cannabis, eds N. Hamel and D. Cross (Berkeley, CA: Ronin Publishing)*, pp. 27–59. Available at: https://books.google.co.za/books?hl=en&lr=&id=1qyh22v5glgC&oi=fnd&pg=PR16&dq=Clarke,+R.C.+(1981).+%E2%80%98The+genetics+and+breeding+of+Cannabis,,+in+Marijuana+Botany:+An+Advanced+Study:+The+Propagation+and+Breeding+of+Distinctive+Cannabis,+eds+N.+Hamel+and+D.+Cross+(Berkeley,+CA:+Ronin+Publishing),+pp.+27%E2%80%9359.+Doi&ots=IA9W9IMTLQ&sig=5HU5d9oSlUVPPHgNIlAETL3zZUY&redir_esc=y#v=onepage&q&f=false (Accessed 22/06/2025)

Clarke, R.C., Merlin, M.D. (2013). '*Cannabis*: Evolution and Ethnobotany', Berkeley, CA: University of California Press. Available at: https://www.ucpress.edu/book/9780520270480/cannabis (Accessed 25/10/2025).

Crawford, S., Rojas, B.M., Crawford, E., Otten, M., Schoenenberger, T.A., Garfinkel, A.R., Chen, H. (2021). 'Characteristics of the diploid, triploid, and tetraploid versions of acannabigerol-dominant F1 hybrid industrial hemp cultivar, *Cannabis sativa* 'stem cell CBG'', *Genes (Basel),* 12(6), pp. 923. Doi: 10.3390/genes12060923.

De Meijer, E.P.M., Hammond, K., Micheler, M. (2009). 'The inheritance of chemical phenotype in *Cannabis sativa* L. (III): variation in cannabichromene proportion', *Euphytica*, 165, pp. 293-311. Doi: 10.1007/s10681-008-9787-1

Doležel, J., Vrána, J., Šimková, H., Burešová, V., Šafář, J. (2014). 'Advances in plant chromosome genomics', *Biotechnology Advances*, 32, pp.122-136. Doi: 10.1016/j.biotechadv.2013.12.011

Casano, S., Grassi, G., Martini, V., Michelozzi, M. (2010). 'Variations in terpene profile of different strains *of Cannabis sativa* L.', *ISHS Acta Horticulture*, 925. Doi: 10.17660/ActaHortic.2011.925.15

Chagné, D., Batley, J., Edwards, D., Forster, J. W. (2007). 'Single nucleotide polymorphisms genotyping in plants', *Plant Genotyping II: SNP Technology*, 53, pp. 77-94. Doi: 10.1007/978-0-387-36011-9_5

Crampton, H. (2015). 'Dagga. A short history (then, now &just now)', *South Africa: Jacana*. Available at: https://books.google.co.za/books/about/Dagga.html?id=XTjqsgEACAAJ&redir_esc=y (Accessed 16/03/2024)

De Filippis, D., Esposite, G., Cirillo, C., Cipriano, M., Winter, B.Y.D., Scuderi, C., Sarnelli, G., Cuoma, R., Steardo, L., Man, J.G., Iuvone, T. (2011). 'Cannabidiol reduces intestinal inflammation through the control of neuroimmune axis', *PloS One*, 6. Doi: 10.1371/journal.pone.0028159

Dewick, P.M. (2009). 'Medicinal natural products: A biosynthetic approach', *3rd ed. John Wiley & Sons: Chichester*. Available at: https://books.google.co.za/books?hl=en&lr=&id=A4zptjOJfKQC&oi=fnd&pg=PA1&dq=Dewick,+P.M.+Medicinal+Natural+Products:+A+Biosynthetic+Approach,+3rd+ed.%3B+John+Wiley+%26+Sons:+Chichester,+UK,+2009&ots=wjdpF-NEYh&sig=wDOcPV0KU6LIyKpOpUoagFQJv8c&redir_esc=y#v=onepage&q&f=false (Accessed 20/03/2024)

DNA Genetics. (2025). 'Durban Poison *Cannabis* strain – Pure South African *sativa*', Available at: https://dnagenetics.com/durban-poison-cannabis-strain/ (Accessed 30/09/2025).

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dala, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., King, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers., K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S. (2009). 'Real-time DNA sequencing from single polymerase molecules', *Science*, 323 (5910), pp.133-138. Doi: 10.1126/science.1162986

Emms, D.M., Kelly, S. (2019). 'OrthoFinder: phylogenetic orthology inference for comparative genomics', *Genome Biol*ogy, 20, 238. Doi: 10.1186/s13059-019-1832-y

Fernandez-Pozo, N., Gao, D., Zhang, J., Li, T., Lin, Z., Mueller, L.A., Tanksley, S.D. (2019). 'Hybrid genome assembly using short and long sequencing reads to improve genetic variation detection', *BMC Genomics*, 20(1), pp.1-14. Doi: 10.1186/s12864-018-5356-8

Felsenstein, J. (1981). 'Evolutionary trees from DNA sequences: A maximum likelihood approach', *Journal of Molecular Evolution***,** 17, pp. 368-376. Doi: 10.1007/BF01734359

Fidan, M., Süzerer, V., Onay, A. (2023). '*Cannabis sativa* L.*:* Origin, distribution, taxonomy and biology'*, Türk Bilimsel Derlemeler Dergisi*, 16(1), pp. 10-28. Available at: https://derleme.gen.tr/index.php/derleme/article/view/454 (Accessed 18/02/2024)

Finola. (2018). 'Some basic things to know about FINOLA', Available at: https://finola.fi/wp-content/uploads/2019/12/Some_basic_things_to_know_about_FINOLA_2018.pdf (Accessed 30/09/2025)

Gagne, S.J., Stout, J.M., Liu, E., Page, J.E. (2012). 'Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides', *Proceedings of the National Academy of Sciences USA*, 109(31), pp. 12811-12816. Doi: 10.1073/pnas.1200330109

Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., Yu, Y., Yang, W., Zhang, Y. (2020). 'A high-quality reference genome of wild *Cannabis sativa*', *Horticulture Research*, 7 (73), pp. 1-11. Doi:10.1038/s41438-020-0295-3.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J., Anderson, L.K. (2007). 'Recombination: an underappreciated factor in the evolution of plant genomes', *Nature Reviews Genetics*, 8, pp.77-84. Doi: 10.1038/nrg1970

Grivennikov, S.I., Greten, F.R., Karin, M. (2010).'Immunity, inflammation, and cancer," *Cell*, 140 (6), pp. 883-899, 2010. Doi: 10.1016/j.cell.2010.01.025

Girgis, H.Z. (2015). 'Red: an intelligent, rapid, accurate tool for detecting repeats *de novo* on the genomic scale', *BMC Bioinformatics*, 16, pp. 227. Doi: 10.1186/s12859-015-0654-5

Gulck, T. & Moller, B.L. (2020). 'Phytocannabinoids: origins and biosynthesis.' *Trends in Plant Science*, 25, pp. 985-1004. Doi: 10.1016/j.tplants.2020.05.005

Goodwin, S., McPherson, J.D., McCombie, W.R. (2016). 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics*, 17(6), pp.333-351. Doi: 10.1038/nrg.2016.49

Grassa, C.J., Wegner, J.P., Dabney, C., Poplawski, S.G., Motley, S.T., Michael, T.P., Schwartz, C.J., Weiblen, G.D. (2018). 'A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content', *bioRxiv*, Doi:10.1101/458083.

Guo, L., Winzer, T., Yang, X., Li, Y., Ning, Z., He, Z., Teodor, R., Lu, Y., Bowser, T.A., Graham, I.A., Ye, K. (2018) 'The opium poppy genome and morphinan production', *Science*, 362, pp. 343-347. Doi: 10.1126/science.aat4096.

Hall, T.A. (1999). 'BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT'*, Nucleic Acids Symposium Series*, 41, pp. 95-98.

Hartsel, J.A., Eades, J., Hickory, B., Makriyannis, A. (2016). '*Cannabis sativa* and hemp', *Nutraceuticals: Efficacy, Safety and Toxicity*, pp. 735-754. Doi: 10.1016/B978-0-12-802147-7.00053-X

Havecker, E.R., Gao, X., Voytas, D.F. (2004) 'The diversity of LTR retrotransposons', *Genome Biology*, 5(6), pp. 225. Doi: 10.1186/gb-2004-5-6-225.

Hawkins, J.S., Proulx, S.R., Rapp, R.R., Wendel, J.F. (2009) 'Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants', *Evolution*, 106(42), pp. 17811–17816. Doi: 10.1073/pnas.0904339106.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P. (2017). 'Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper', *Molecular Biology and Evolution*, 34(8), pp.2115-2122. Doi: 10.1093/molbev/msx148

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. (2018). 'UFBoot2: Improving the ultrafast bootstrap approximation', *Molecular Biology and Evolution*, 35(2), pp.518–522. Doi:10.1093/molbev/msx281

Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C. (2009). 'Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes', *Genome Research*, 19, pp.1270-1278. Doi: 10.1101/gr.088633.108.

Hourfane, S., Ali, S.A., Farag, M.A., Ezzat, S.M. (2023) 'A comprehensive review on *Cannabis sativa* L.: Phytochemistry, pharmacology, and therapeutic potential', *Plants*, 12(6), pp. 1245. Doi: 10.3390/plants12061245

Huang, A.C., Jiang, T., Liu, Y.-X., Bai, Y.-C., Reed, J., Qu, B., Goossens, A., Nützmann, H.-W., Bai, Y., Osbourn, A. (2019) 'A specialized metabolic network selectively modulates *Arabidopsis* root microbiota', *Science*, 364, eaau6389. Doi: 10.1126/science.aau6389.

Innes, P.A., Vergara, D. (2023). 'Genomic description of critical cannabinoid biosynthesis genes', *Botany*, 101(7), pp. 270-283. doi: 10.1139/cjb-2022-0140

Jansen, C., Shimoda, L.M.N., Kawakami, J.K., Ang, L., Bacani, A.J., Baker, J.D., Badowski, C., Lewis, R.J., Manning, J.J., Small-Howard, A.L., Turner, H. (2019). 'Myrcene and terpene regulation of TRPV1', *Channels*, 13(1), pp. 344-366. Doi: 10.1080/19336950.2019.1654347.

Jensen, P.R. (2016). 'Natural products and the gene cluster revolution', *Trends in Microbiology*, 24 (12), pp. 968-977. Doi: 10.1016/j.tim.2016.07.006

Jin, D., Henry, P., Shan, J., Chen, J. (2021). 'Identification of chemotypic markers in three *Cannabis* chemotypes', *Frontiers in Plant Science*, 12, pp. 699530. Doi: 10.3389/fpls.2021.699530

Johnston, H.R., Lee, S., Kim, Y. (2015). 'Population genetics identifies challenges in analyzing next-generation sequencing data', *Human Genetics*, 134(7), pp. 723–734. Doi: 10.1007/s00439-015-1616-3

Johnson, M.S., Wallace, J.G. (2021). 'Genomic and chemical diversity of commercially available high-CBD industrial hemp accessions', *Frontiers in Genetics*, 12, pp. 682475. Doi: 10.3389/fgene.2021.682475

Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A., Medema, M.H. (2017). 'plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters', *Nucleic Acids Research*, 45(W1), pp. W55-W63.Doi: 10.1093/nar/gkx305

Khalsa, J.H., Bunt, G., Blum, K., Maggirwar, S.B., Galanter, M., Potenza, M.N. (2022). 'Review: Cannabinoids as medicinals', *Current Addiction Reports*, 9, pp. 630–646. Doi: 10.1007/s40429-022-00438-3

Kojoma, M., Seki, H., Yoshida, S., Muranaka, T. (2006). 'DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L.', *Forensic Science International*, 159 (2-3), pp. 132–140. Doi: 10.1016/j.forsciint.2005.07.005

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), pp. 1547-1549. Doi: 10.1093/molbev/msy096

Kurtz, L.E., Brand, M.H., Lubell-Brand, J.D. (2020). 'Production of tetraploid and triploid hemp', *HortScience*, 55, pp. 1703-1707. Doi: 10.21273/HORTSCI15303-20

Lanciano, S., Mirouze, M. (2017) 'Transposable element driven innovation in plants', *Current Opinion in Plant Biology*, 36, pp. 18-26. Doi: 10.1016/j.gde.2018.04.002

Laverty, K.U., Stout, J.M., Sullivan, M.J., Shah, H., Gill, N. (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. Doi: 10.1101/gr.242594.118.

Lewin, H.A., Larkin, D.M., Pontius, J., O'Brien, S.J. (2009). 'Every genome sequence needs a good map', *Genome Research*, 19, pp.1925-1928. Doi: 10.1101/gr.094557.109

Lewin H.A., Richards S., Lieberman Aiden E., Allende M.L., Archibald J.M., Bálint M., Barker K.B., Baumgartner B., Belov K., Bertorelle G., Blaxter, M.L., Cai, J., Caperello, N.D., Carlson, K., Castilla-Rubio, J.C., Chaw, S., Chen, L., Childers, A.K., Coddington, J.A., Conde, D.A., Corominas, M., Crandall, K.A., Crawford, A.J., Durbn, R., Ebenezer, T.E., Edwards, S.V., Fedrigo, O., Flicek, P., Formenti, G., Gibbs, R.A., Thomas, M., Gilbert, P., Goldstein, M.M., Graves, J.M., Greely, H.T., Grigoriev, I.V., Hackett, K.J., Hall, N., Haussler, D., Helgen, K.M., Hogg, C., Isobe, S., Jakobsen, K.S., Janke, A., Jarvis, E.D., Johnson, W.E., Jones, S.J.M., Karlsson, E.K., Kersey, P.J., Kim, J.-H., Kress, W.J., Kuraku, S., Lawniczak, M.K.N., Leebens-Mack, J.H., Li, X., Lindblad-Toh, K., Liu, X., Lopez, J.V., Marques-Bonet, T., Mazard, S., Mazet, J.A.K., Mazzoni, C.J., Myers, E.W., O'Neill, R.J., Paez, S., Park, H., Robinson, G.E., Roquet, C., Ryder, O.A., Sabir, J.S.M., Shaffer, H.B., Shank, T.M., Sherkow, J.S., Soltis, P.S., Tang, B., Tedersoo, L., Uliano-Silva, M., Wang, K., Wei, X., Wetzer, R., Wilson, J.L., Xu, X., Yang, H., Yoder, A.D. and Zhang, G. (2022). 'The Earth BioGenome Project: Starting the clock', *Proceedings of the National Academy of Sciences*, 119, article 4. Doi: 10.1073/pnas.2115635118

Li, H., Durbin, R. (2009). 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*, 25(14), pp.1754-1760. Doi:10.1093/bioinformatics/btp324

Li, H., Ruan, J., Durbin, R. (2010). 'Mapping short DNA sequencing reads using the Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp.1754-1760. Doi: 10.1093/bioinformatics/btp698

López-Giráldez, F., Townsend, J. P. (2011). 'PhyDesign: an online application for profiling phylogenetic informativeness', *BMC Evolutionary Biology*, 11, pp. 152. Doi: 10.1186/1471-2148-11-152

Lowe, T. M. & Chan, P. P. (2016). 'tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes', *Nucleic Acids Res*earch, 44, pp. W54–W57. Doi: 10.1093/nar/gkw413

Louveau, T., Osbourn, A. (2019). 'The Sweet Side of Plant-Specialized Metabolism', *Cold Spring Harb Perspect Biology*, ;11(12), a034744. Doi: 10.1101/cshperspect.a034744.

Luo, X., Luo, X., Reiter, M.A., d'Espaux, L., Wong, J., Denby, C.M., Lechner, A., Zhang, Y., Grzybowski, A.T., Harth, S., Lin, W., Lee, H., Yu, C., Shin, J., Deng, K., Benites, V.T., Wang, G., Baidoo, E.E.K., Chen, Y., Dev, I., Petzold, C.J., Keasling, J.D. (2019). 'Complete biosynthesis of cannabinoids and their unnatural analogues in yeast', *Nature*, 567, pp 123-126. Doi: doi.org/10.1038/s41586-019-0978-9

Lynch, R.C., Vergara, D., Tittes, S., White, K., Schwartz, C.J., Gibbs, M.J., Ruthenburg, T.C., DeCesare, K., Land, D.P., Kane, N.C. (2016). 'Genomic and chemical diversity in *Cannabis*'. *Critical reviews in plant sciences,* 35 (5-6), pp. 349-363. Doi: 10.1080/07352689.2016.1265363

Malicki, M., Malik, H.S., Eickbush, T.H. (2020). 'DIRS retrotransposons amplify via linear, single-stranded intermediates', *Nucleic Acids Research*, 48(8), pp. 4230-4242. Available at: https://academic.oup.com/nar/article/48/8/4230/5805300 (Accessed 17/10/2025)

Marks, M.D., Tian, L., Wenger, J.P., Omburo, S.N., Soto-Fuentes, W., Gang, D.R., Weiblen, G.D., Dixon, R.A. (2009). 'Identification of candidate genes affecting delta-(9)-tetrahydrocannabinol biosynthesis in *Cannabis sativa'*, *Journal of Experimental Bot*ony, 60, pp. 3715-3726. Doi: 10.1093/jxb/erp210

Martin, R., Hackl, T., Hattab, G., Fischer, M.G., Heider, D. (2020). 'MOSGA: modular open-source genome annotator', *Bioinformatics, 3*6(22–23), pp.5514–5515. Doi: 10.1093/bioinformatics/btaa1003

McKernan, K.J., Helbert, Y., Kane, L.T., Ebling, H., Zhang, L., Liu, B., Eaton, Z., McLaughlin, S., Kingan, S., Baybayan, P., Concepcion, G., Jordan, M., Riva, A., Barbazuk, B., Harkins, T. (2020). 'Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes', *bioRxiv*. Doi:10.1101/2020.01.03.894428

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R. (2020). 'IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era', *Molecular Biology and Evolution*, 37(5), pp.1530–1534. Doi: 10.1093/molbev/msaa015

Mirny, L.A., Gelfand, M.S. (2002). 'Using orthologous and paralogous proteins to identify specificity determining residues', *Genome Biol*ogy, 3. Doi: 10.1186/gb-2002-3-3-preprint0002

Miyazawa, M., Yamafuji, C., Sumire, K., Nakamura, S. (2011) 'Inhibition of acetylcholinesterase activity by bicyclic monoterpenoids', *Journal of Agricultural and Food Chemistry*, 59(10), pp. 5372-5376. Doi: 10.1021/jf200661z.

Mudge, E.M., Murch, S.J., Brown, P.N. (2018). 'Chemometric analysis of cannabinoids: Chemotaxonomy and metabolomics', *Scientific Reports*, 8(1), pp.13190. Doi: 10.1038/s41598-018-31120-2

Notredame, C., Higgins, D.G., Heringa, J. (2000). 'T-coffee: a novel method for fast multiple sequence alignment', *Journal of Molecular Biology*, 302(1), pp. 205-217. Doi: 10.1006/jmbi.2000.4042

Pacifico, D., Miselli, F., Micheler, M., Carboni, A., Ranalli, P., Mandolino, G. (2006). 'Genetics and marker-assisted selection of the chemotype in *Cannabis sativa*', *Euphytica*, 148(1-2), pp.1-12. Doi: 10.1007/s11032-005-5681-x

Parsons, J.L., Martin, S.L., James, T., Golenia, G., Boudko, E.A., Hepworth, S.R. (2019). 'Polyploidization for the genetic improvement of *Cannabis sativa'*, Frontiers in Plant Science, 10, pp. 476. Doi: 10.3389/fpls.2019.00476.

Paterson, A.H., Bowers, J.E., Burow, M.D., Draye, X., Elsik, C.G., Jiang, C.-X., Katsar, C.S., Lan, T.-H., Lin, Y.-R., Ming, R., Wright, R.J. (2000). 'Comparative genomics of plant chromosomes', *The Plant Cell*, 12(9), pp.1523-1539. Doi: 10.1105/tpc.12.9.1523

Petrovic, G., Aleixandre-Tudo, J.L., Buica, A. (2019). 'Grape must profiling and cultivar discrimination based on amino acid composition and general discriminant analysis with best subset', *South African Journal of Enology and Viticulture*, 40, article 2. Doi: 10.21548/40-2-3373

Piednoël, M., Gonçalves, I.R., Higuet, D., Bonnivard, E. (2011). 'Eukaryote DIRS1-like retrotransposons: an overview', *BMC Genomics*, 12(1), p. 621. Available at: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-621 (Accessed 17/10/2025)

Polturak, G., Osbourn, A. (2021) 'The emerging role of biosynthetic gene clusters in plant metabolism', *New Phytologist*, 230(5), pp. 1535-1546. Doi: 10.1111/nph.17250.

Raina, M. (2014) 'tRNAs as regulators of biological processes', *Frontiers in Genetics*, 5, pp. 171. Doi: 10.3389/fgene.2014.00171.

Rea, G., Metoui, O., Infantino, A., Federico, R., Angelini, R. (2002). 'Copper amine oxidase expression in defense responses to wounding and *Ascochyta rabiei* invasion', *Plant Physiology*, 128(3), pp. 865-875. Doi: 10.1104/pp.010646.

Ren, G., Zhang, X., Li, X., Wang, X., Zhang, J., Zhang, H., Wang, Y., Li, X. (2021). 'Large-scale whole-genome resequencing unravels the genetic architecture of cannabinoid biosynthesis in *Cannabis sativa*', *Science Advances*, 7(33), pp. eabg2286. Doi: 10.1126/sciadv.abg2286

Rotherham, D. & Harbison, S.A. (2010). 'Differentiation of drug and non-drug *Cannabis* using a single nucleotide polymorphism (SNP) assay', *Forensic Science International*, 207, pp. 193-197. Doi: 10.1016/j.forsciint.2010.10.006

Russo, E.B. (2011) 'Taming THC: Potential *Cannabis* synergy and phytocannabinoid-terpenoid entourage effects', *British Journal of Pharmacology*, 163(7), pp. 1344-1364. Doi: 10.1111/j.1476-5381.2011.01238.x.

Ryu, B.R., Gim, G.J., Shin, Y.R., Kang, M.J., Kim, M.J., Kwon,. T.H,. Lim, Y.S., Park, S.H., Lim, J.D. (2024). 'Chromosome-level haploid assembly of *Cannabis sativa* L. cv. Pink Pepper', *Scientific Data*, 11, pp. 1-9. Doi:10.1038/s41597-024-04288-8

Saitou, N., Nei, M. (1987). 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular Biology and Evolution*, 4(4), pp. 406-425. Doi: 10.1093/oxfordjournals.molbev.a040454

de Sousa, D.P., Raphael, E., Brocksom, U., Brocksom, T.J. (2007) 'Sedative effect of monoterpene alcohols in mice: a structure–activity relationship study', *Zeitschrift für Naturforschung C*, 62(5-6), pp. 563-568. Doi: 10.1515/znc-2007-5-620.

Stott, C.M., Bobay, L.-M. (2020). 'Impact of homologous recombination on core genome phylogenies', *BMC Genomics*, 21, pp. 829. Doi: 10.1186/s12864-020-07262-x

Seemann, T. (2018). 'barrnap 0.9: rapid ribosomal RNA prediction [Internet]' *GitHub repository*, Available from: https://github.com/tseemann/barrnap.

Shoji, T., Kajikawa, M., Hashimoto, T. (2010) 'Clustered transcription factor genes regulate nicotine biosynthesis in tobacco', *The Plant Cell*, 22, pp. 3390-3409. Doi: 10.1105/tpc.110.078543.

Smit, A. F. A., Hubley, R., Green, P. (2013). *RepeatMasker Open-4.0*. Available at: http://www.repeatmasker.org/RMDownload.html (Accessed 08/04/2024)

Smit, S.J., Wurtzel, O., Osbourn, A. (2022) 'Plant biosynthetic gene clusters in the context of metabolic evolution', *Natural Product Reports*, 39, pp. 374-392. Doi: 10.1039/d2np00005a.

Tan, Z.G., Clomburg, J.M., Conzalez, R. (2018). 'Synthetic pathway for the production of olivetolic acid in *Escherichia coli'*, *ACS Synthetic Biolpgy*, 7(8). Doi: 10.1021/acssynbio.8b00075

Taura F, Morimoto S, Shoyama Y. (1996). 'Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa* L. biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid', *Journal of Biological Chem*istry, 271, pp 17411-17416. Doi: 10.1077/jbc.271.29.17411

Taura, F., Sirikantaramas, S., Shoyama, Y.Y., Yoshikai, K., Morimoto, S. (2007). 'Cannabidiolic acid synthase, the chemotype-determining enzyme in the fiber-type *Cannabis sativa*', *FEBS Letters*, 581(16), pp.2929-2934. Doi: 10.1016/j.febslet.2007.05.043

Tian, S., Wang, D., Yang, L., Zhang, Z., Liu, Y. (2022) 'A systematic review of 1-Deoxy-D-xylulose-5-phosphate synthase in terpenoid biosynthesis in plants', *Plant Growth Regulation*, 96, pp. 221-235. Doi: 10.1007/s10725-021-00784-8.

Todaro, B. (2012). 'Cannabinoids in the treatment of chemotherapy induced nausea and vomiting', *Journal of the National Comprehensive Cancer Network*, 10, pp.487-492. Doi: 10.6004/jnccn.2012.0048

Tomas, B.F, ElSohly, M.A. (2015). 'The analytical chemistry of *Cannabis:* quality assessment, assurance and regulation of medicinal marijuana and cannabinoid preparations', *Elsevier*

Amsterdam1st edition. Available at: https://www.amazon.com/Analytical-Chemistry-Cannabis-Cannabinoid-Preparations/dp/0128046465 (Accessed 13/03/2024)

University of Toronto. (2018). '*Cannabis sativa* Purple Kush genome assembly ASM23057v5', Available at: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_000230575.5 (Accessed 30/09/2025).

Van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., Page, J.E. (2011). 'The draft genome and transcriptome of *Cannabis sativa*', *Genome Biology,* 12: R102. Doi: 10.1186/gb-2011-12-10-r102

Vergara, D. (2019). 'Gene copy number is associated with phytochemistry in *Cannabis sativa*', *AoB Plants,* 11(6), pp. lz074. Doi: 10.1093/aobpla/plz074

Vitte, C., Panaud, O. (2005). 'LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model', *Cytogenetic and Genome Research*, 110(1-4), pp.91-107. Doi: 10.1159/000084941

Wang, S., Zhang, Y., Zhang, X., Li, D., Liu, Y., Li, X., Zhang, X., Zhang, L., Zhang, X. (2025). 'Pan-genome analysis of *Cannabis sativa*', *International Journal of Molecular Sciences*, 26(17), pp. 8354. Doi: 10.3390/ijms26178354

Welling, M.T. (2020). 'An extreme-phenotype genome-wide association study identifies loci associated with cannabinoid content in *Cannabis sativa*', *Nature Plants*, 6(6), pp.601-608. Doi: 10.1038/s41477-020-0666-3

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H. (2007) 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics*, 8(12), pp. 973-982. Doi: 10.1038/nrg2165.

Winkler, A., Lyskowski, A., Riedl, S., Puhl, M., Kutchan, T.M., Macheroux, P., Gruber, K. (2008). 'A concerted mechanism for berberine bridge enzyme', *Nature Chemical Biology*, 4(12), pp. 739-741. Doi: 10.1038/nchembio.123.

Wong, T.K.F., Ly-Trong, N., Ren, H., Banos, H., Roger, A.J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M.W., Huttley, G., Lanfear, R. & Minh, B.Q. (2025). '*IQ*-TREE 3: Phylogenomic inference software using complex evolutionary models', *Ecoevorxiv*, Available at: https://ecoevorxiv.org/repository/view/8916/ (Accessed 11/09/2025).

Xie, Z., Mi, Y., Kong, L., Gao, M., Chen, S., Chen, W., Meng, X., Wei Sun, W., Shilin Chen, S., Xu, Z. (2023). '*Cannabis sativa*: origin and history, glandular trichome development, and cannabinoid biosynthesis', *Horticulture Research,* 10. Doi: 10.1093/hr/uhad150

Yao, X., Nie, J., Bai, R., Sui, X. (2020). 'Amino acid transporters in plants: Identification and function', *Plants (Basel*), 9(8), pp. 972. doi: 10.3390/plants9080972

You, F.M. (2023). 'Plant genomics-Advancing our understanding of plants', *International Journal of Molecular Sciences*, 24(14), p.11528. Doi: 10.3390/ijms241411528

Zager, J.J., Lange, I., Srividya, N., Smith, A., Lange, B.M. (2019). 'Gene networks underlying cannabinoid and terpenoid biosynthesis in *Cannabis sativa*', *Frontiers in Plant Science*, 10, pp. 1422. Doi: 10.3389/fpls.2019.01422

Zuardi, A.W. (2006). 'History of *cannabis* as a medicine: A review', *Brazilian Journal of Psychiatry*, 28(2), pp. 153–157. Doi: 10.1590/S1516-44462006000200015

Chapter Four: Summary

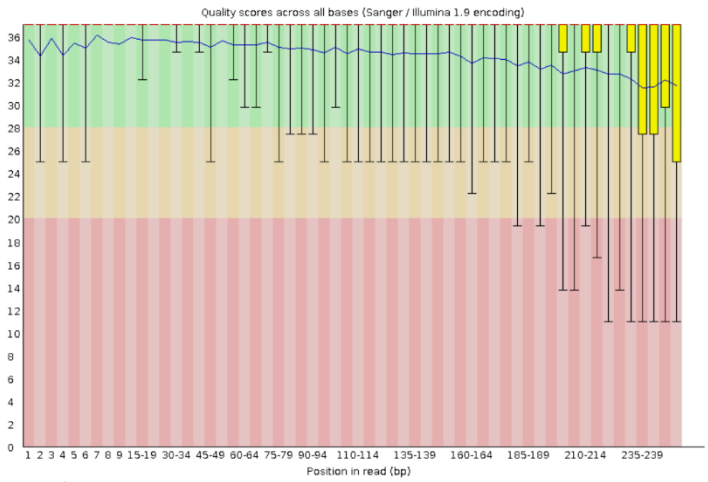
Cannabis sativa L. produces 125 cannabinoids with the four most abundant being: delta-9-tetrahydrocannabinol (Δ^9-THC), cannabidiol (CBD), cannabichromene (CBC) and cannabigerol (CBG). Despite the growing number of *C. sativa* genome assemblies, high-quality references for African landraces remain absent. To address this gap, this study focused on the African landrace *C. sativa* Swazi Gold (SG), originating from the Kingdom of Eswatini, with a known drug-type chemotype and cultivation for its potent psychoactive effects and Δ^9-THC levels (16-18%). The genome of *C. sativa* SG contains important information required to understand the biosynthesis pathways of the plant and its cannabinoids. In this study, we have begun to address this genomic research gap surrounding African landraces. Through the genome assembly, annotation and comparative analysis we can explore the genomics behind *C. sativa's* diverse applications.

The draft genome for *C. sativa* SG was assembled using a hybrid *de novo* approach (Chapter 2), which combined both low coverage long-read (Pacific Bioscience) and high coverage short-read (Illumina NovaSeq 6000) sequencing data. Multiple *de novo* assembly tools, combined with reference-guided assemblers and refining tools were employed for genome assembly, followed by structural and functional annotation. The final draft *C. sativa* SG genome had a total size of 764.62 Mb, G+C content of 34.04% and a completeness score of 97.0%. A total of 45,142 protein coding genes, 648 tRNAs and 907 rRNAs were predicted within the assembly, while repetitive elements encompassed a predicted 77.19% of the genome. These values fall within the same ranges achieved for the *C. sativa* genomes available on the NCBI database, with high similarity to the reference genome (Pink Pepper), suggesting a high level of accuracy and reliability in the achieved draft genome for *C. sativa* SG.

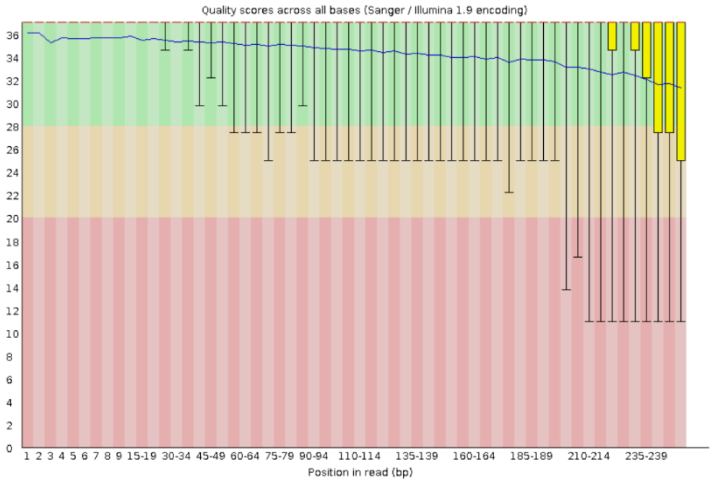
The *C. sativa* SG draft genome was compared against the genomes of 14 *C. sativa* landraces and cultivars, to elucidate genomic differences in the cannabinoid profiles (Chapter 3). BLAST analysis indicated the presence of 10/11 cannabinoid biosynthesis genes within the draft *C. sativa* SG genome, confirming its potential to produce all the major cannabinoids. Core genome phylogenetic analysis resulted in mixed clustering of the different chemotypes, while analysis of cannabinoid biosynthetic enzymes evidence of extensive paralogy for most enzymes. This is consistent with the cross-hybridization and selective-breeding of *C. sativa* cultivars for their desirable traits. Additional biosynthetic gene clusters (BGCs) analysis revealed the most abundant BGC types to be saccharides, alkaloids and terpenes among the genomes, with many putative clusters in SG which suggests the potential for novel BGCs.

Overall, comparative genomic analysis revealed the challenges of evaluating evolutionary relationships and chemotype classification in *C. sativa*, due to extensive laboratory-driven and selective crossbreeding to produce specific chemical profiles. Instead, the results highlight the significant influence that decades of domestication and desired trait selection, through artificial breeding, have had on the evolution of *C. sativa*. Further, this study brought to the fore the power of genomics in elucidating these complexities. Future research should focus on exploring the regulatory mechanisms and gene copy number variations (CNVs) that shape chemotype expression, conducting comparative genome analyses into cannabinoid biosynthetic gene expression patterns across chemotypes and validating biosynthetic gene variants and their evolutionary impact.

Chapter 5: Appendix

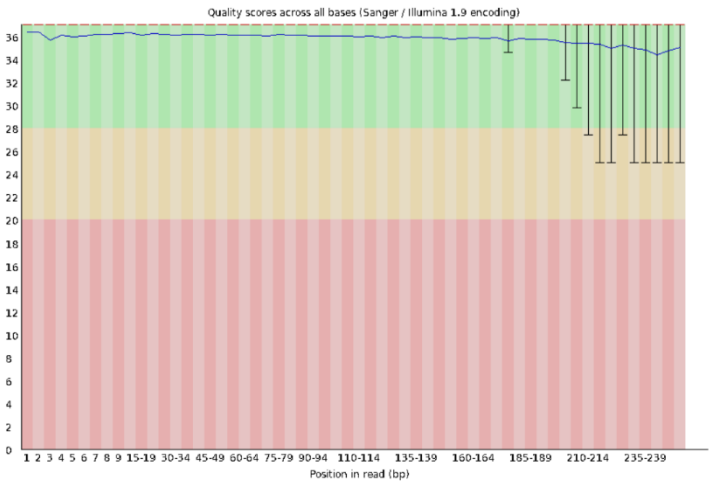


R1

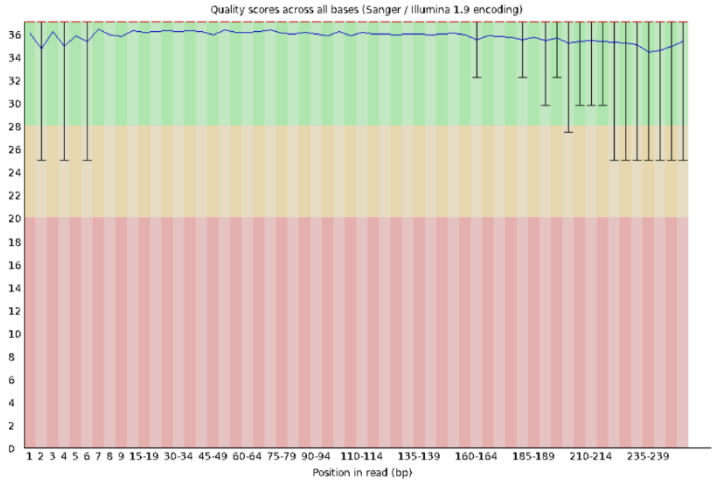


R2

Figure A1: Quality of the *Cannabis sativa* Swazi Gold raw Illumina short-read data.



R1



R2

Figure A2: Quality of the *Cannabis sativa* Swazi Gold Illumina short-read data post trimming using TrimGalore v. 0.6.7 with a Phred score cutoff of Q20.

Table A1: *Cannabis sativa* genome metrics for thirteen genomes retrieved from the NCBI database for different cultivars and landraces.

Cultivar / Landrace	Accession	Genomes size (Mb)	Ungapped length (Mb)	Scaffold N50 (Mb)	Contig N50 (Mb)	No. of scaffolds	No. of contigs	G+C content (%)	Genome coverage (×)
Pink Pepper (ref)	GCA_029168945.1	770.3	770.3	77	23.5	17	70	34	299
Abacus (Csat)	GCA_025323715.1	796.6	796.5	80.6	3.2	160	1,023	34	83
Cannbio-2	GCA_016165845.1	914.4	913.5	91.5	172.4 kb	147	8,919	34	86
Cherry Pie	GCA_049724855.1	802.2	752.6	81.1	1.6	721	1,882	34.5	18
cs10	GCF_900626175.2	875.7	736.2	91.9	2	220	1,051	33.5	100
Finola	GCA_003417725.2	1,000	1,000	370.5 kb	370.5 kb	5,303	5,303	34	97.64
JL	GCA_013030365.1	812.5	811.8	83	510 kb	483	2,978	34	153
Jamaican Lion Dash (JL Dash)	GCA_003660325.2	1,100	1,100	3.8	3.8	556	556	34	125
Jamaican Lion ^4 (JL5)	GCA_016415525.1	999.1	999.1	3.5	3.5	658	658	34	85
Jamaican Lion ^4 (JL Father)	GCA_013030025.1	1,000	1,000	1.7	1.7	1,263	1,263	34	125
Jamaican Lion ^4 (JL Mother)	GCA_012923435.1	876.6	876.6	3.3	3.3	1,444	1,444	34	125
Punto Rojo	GCA_049725155.1	811.3	758.1	87.6	2	469	1,325	34	18
Purple Kush	GCA_000230575.5	892	891.3	133.9 kb	133.9 kb	12,836	12,836	34	79

Table A2: Annotation results for the fifteen *Cannabis sativa* genomes used in this study.

Cultivar / Landrace	Genome size (Mb)	No. of rRNAs	No. of tRNAs	RED repeats (%)	Repeatmasker repeats (%)	BUSCO completeness (%)	No. of protein-coding genes	SwissProt protein hits	Augustus		EggNOGG protein hits
									No. of genes	No. of introns	
Pink Pepper (ref)	770.69	1085	657	62.77	78.51	97.7	44,396	59,505	45,808	142,531	35,428
Abacus (Csat)	796.58	625	725	60.82	76.63	97.1	51,232	64,313	52,859	158,374	39,090
Cannbio-2	914.40	1168	1156	58.69	74.52	97.1	56,772	60,681	58,657	195,057	46,287
Cherry Pie	802.24	3838	703	58.95	73.34	94.0	43,575	57,884	44,860	137,409	34,456
cs10	876.20	985	602	50.18	64.92	94.1	45,940	58,609	47,324	146,525	34,607
Durban Poison (DP)	748.82	847	648	59.09	76.44	97.3	42,723	59,445	44,092	138,996	34,027
Finola	1009.67	837	861	63.47	79.12	95.7	56,978	62,228	58,979	181,443	45,357
JL	812.53	758	689	60.39	76.86	92.7	48,819	59,186	50,429	156,756	39,068
Jamaican Lion Dash (JL Dash)	1333.38	1208	1355	67.24	77.47	98.2	58,966	59,664	61,156	208,679	47,605
Jamaican Lion ^4 (JL5)	999.12	1367	857	62.87	77.96	98.0	56,485	60,910	58,325	184,333	45,288
Jamaican Lion ^4 (JL Father)	1009.16	685	845	62.85	78.25	98.0	56,582	60,503	58,464	185,262	45,509
Jamaican Lion ^4 (JL Mother)	876.74	856	710	63.65	79.16	97.5	49,064	59,975	50,068	156,474	39,013
Purple Kush	891.96	830	1077	59.92	75.76	93.5	54,649	60,332	56,512	181,746	44,285
Punto Rojo	811.90	1200	757	58.01	72.72	97.9	44,357	59,472	45,720	142,550	35,399
Swazi Gold	764.62	907	648	59.24	77.19	97.0	43,646	10,206	45,142	140,786	34,950

Table A3: Comparison of the composition of repetitive elements in the Rebaler reference-guided *Cannabis sativa* Swazi Gold genome to the composition repeat element range for the fourteen annotated *C. sativa* genomes, determined using RepeatMasker.

Transposable element	**Type**	**Swazi Gold genome total elements**	**Swazi Gold genome elements (%)**	**14 *Cannabis sativa* genomes total elements range**	**Genomic proportion (%) of transposable elements across 14 *Cannabis sativa* genomes**
Retrotransposons	SINEs	390	0.00	362-693	0.00
	LINEs	27,965	3.21	27,185-51,372	2.80-3.70
	L1/CIN4	27,965	3.21	27,185-51,372	2.80-3.70
	LTR elements	187,047	47.63	182,191-319,968	39.97-49.65
	Ty1/Copia	111,111	27.31	108,645-190,968	23.57-28.53
	Gypsy/DIRS1	75,727	20.27	73,343-128,639	16.36-21.28
DNA transposons	hobo-Activator	4,925	1.34	4,157-7,634	0.62-1.78
	MULE-MuDR	17,177	1.91	16,156- 30,763	1.64-2.02
	Tourist /Harbinger	1,975	0.09	1,930- 3,676	0.08-0.11
Rolling-circles	Rolling-circles	938	0.13	878-1,666	0.10-0.15
Unclassified	Unclassified	356 908	21.81	347,852-645,160	18.60-22.56

Table A4: The rRNA predications using Barrnap v. 0.9. for the fifteen *Cannabis sativa* genomes used in this study.

Cultivar / Landrace	28S rRNA	18S rRNA	5S rRNAs	Total rRNAs
Pink Pepper (ref.)	10	6	1069	1085
Abacus (Csat)	9	8	606	625
Cannbio-2	184	160	744	1168
Cherry Pie	463	429	2544	3838
cs10	87	85	754	985
Durban Poison (DP)	12	6	829	847
Finola	22	23	783	837
JL	6	3	745	758
Jamaican Lion Dash (JL Dash)	115	104	900	1208
Jamaican Lion ^4 (JL5)	205	194	784	1367
Jamaican Lion ^4 (JL Father)	8	6	666	685
Jamaican Lion ^4 (JL Mother)	16	12	817	856
Purple Kush	102	60	644	830
Punto Rojo	38	33	1112	1200
Swazi Gold	12	6	899	907

Table A5: Local BLAST results for the identification of cannabinoid biosynthetic genes in the annotated *Cannabis sativa* genomes. AP: aromatic prenyltransferase; ACCase: acetyl-CoA carboxylase; CBDAS: cannabidiolic acid synthase; CBCAS: cannabichromenic acid synthase; DXS: 1-deoxy-D-xylulose-5-phosphate synthase; HCAS; hexanoyl-CoA synthetase; GPPS: geranyl pyrophosphate synthase; GOT: geranylpyrophosphate olivetolate geranyltransferase; OAC: olivetolic acid cyclase; OLS: olivetol synthase; THCAS: tetrahydrocannabinolic acid synthase.

Cultivar / Landrace	Proteins not found	Proteins found
Pink Pepper (ref.)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Abacus (Csat)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Cannbio-2	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Cherry Pie	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
cs10	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Durban Poison (DP)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Finola	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
JL	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Jamaican Lion Dash (JL Dash)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Jamaican Lion ^4 (JL5)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Jamaican Lion ^4 (JL Father)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Jamaican Lion ^4 (JL Mother)	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Purple Kush	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Punto Rojo	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS
Swazi Gold	AP	ACCase, CBDAS, CBCAS, DXS, HCAS, GPPS, GOT, OAC, OLS, THCAS

Table A6: Biosynthetic gene cluster (BGC) identification results for all fifteen annotated *Cannabis sativa* genomes using PlantiSMASH (Kautsar *et al.,* 2017).

Cultivar / Landrace	Total clusters	Saccharide	Putative	Terpene	Lignan-Fatty acid	Phenolamide-Alkaloid	Alkaloid	Saccharide-Terpene/ Saccharide-	Fatty acid-Saccharide/	Total clusters	Saccharide	Putative	Terpene	Lignan-Fatty acid	Phenolamide-Alkaloid
Pink Pepper (ref.)	31	9	7	3	1	0	5	1	1	1	0	0	1	0	2
Abacus (Csat)	33	10	5	5	1	1	5	1	1	0	1	1	1	0	0
Cannbio-2	36	10	8	6	1	1	4	1	1	0	0	2	1	1	0
Cherry Pie															
cs10	35	11	8	6	1	0	5	1	1	1	0	1	0	0	0
Durban Poison (DP)	33	10	7	4	1	0	6	1	1	1	0	1	1	0	0
Finola	42	15	9	3	1	1	5	2	1	0	1	1	1	1	1
JL	33	5	10	3	1	1	6	2	2	0	0	1	1	0	1
Jamaican Lion Dash	43	14	8	5	1	1	7	0	1	0	0	2	2	0	2
Jamaican Lion ^4 (JL5)	47	14	8	6	2	1	8	1	1	1	0	2	2	0	1
Jamaican Lion ^4 (JL Father)	41	12	6	5	2	1	7	2	2	2	0	1	1	0	0
Jamaican Lion ^4 (JL Mother)	27	8	4	2	1	1	5	3	1	0	0	1	1	0	0
Punto Rojo	27	10	6	2	1	1	4	1	1	1	0	0	0	0	0
Purple Kush	42	12	11	4	0	1	4	3	2	1	2	1	1	0	0
Swazi Gold	31	11	7	3	1	0	5	1	1	1	0	0	0	0	1