(12) UK Patent Application (19) GB (11) 2617110 (13) A

(43) Date of A Publication 04.10.2023

(21) Application No: 2204468.9

(22) Date of Filing: 29.03.2022

(71) Applicant(s):
Puregene AG
Etzmatt 273, Zeiningen 4314, Switzerland

(72) Inventor(s):
Maximilian Moritz Vogt
Claudio Cropano
Dániel Árpád Carrera
Gavin Mager George
Michael Eduard Ruckle
Leron Katsir
Mercedes Thieme
Michele Wyler

(74) Agent and/or Address for Service:
WP Thompson
138 Fetter Lane, LONDON, EC4A 1BT,
United Kingdom

(51) INT CL:
C12Q 1/6827 (2018.01) ***A01H 1/04*** (2006.01)
A01H 6/28 (2018.01) ***C07K 14/415*** (2006.01)
C12N 15/82 (2006.01) ***C12Q 1/68*** (2018.01)

(56) Documents Cited:
WO 2020/093101 A1
Genome Biol., Vol.12, 2011, van Bakel, H. et al., "The draft genome and transcriptome...", p.R102
J. Cannabis Res., Vol.1, 2019, Schwabe, A. L. & McGlaughlin, M. E., "Genetic tools weed out misconceptions...", Article No.: 3

(58) Field of Search:
Other: **WPI, EPODOC, Patent Fulltext, BIOSIS, MEDLINE**

(54) Title of the Invention: **Quantitative trait loci associated with purple color in cannabis**
Abstract Title: **Quantitative trait loci associated with purple colour in cannabis**

(57) A method of identifying a *Cannabis sativa* plant comprising quantitative trait loci (QTLs) associated with purple color, and to Cannabis sativa plants comprising the QTLs. The method comprises genotyping a plant to identify polymorphisms (SNPs) associated with purple colour. The invention further relates to marker assisted selection and marker assisted breeding methods for obtaining plants having purple color, as well as to methods of producing Cannabis sativa plants with the absence of purple color and/or varying degrees of purple color and plants.

GB 2617110 A

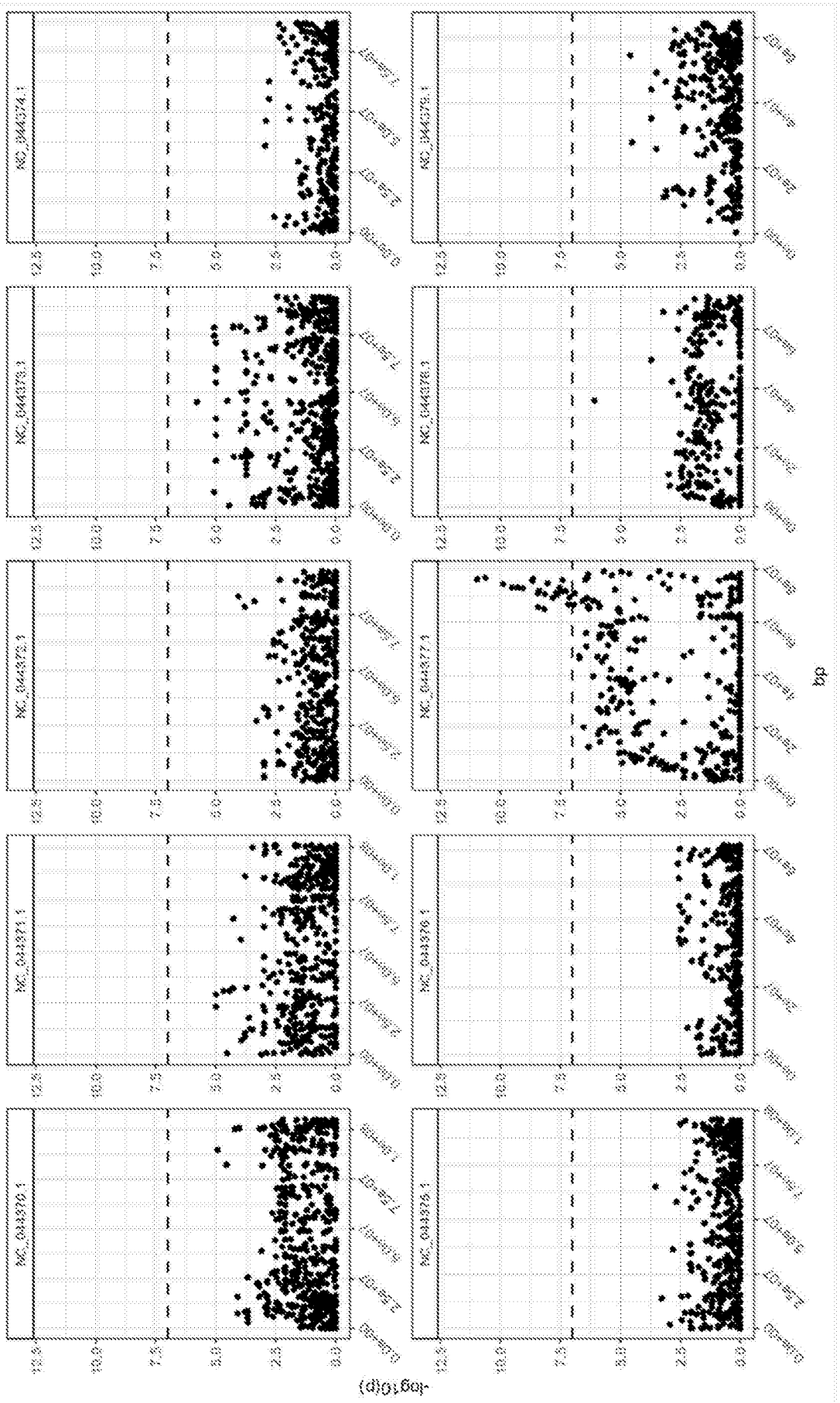
NC_044374.1
NC_044375.1
NC_044376.1
NC_044377.1
NC_044378.1
NC_044379.1
-log10(p)
bp

Figure 1

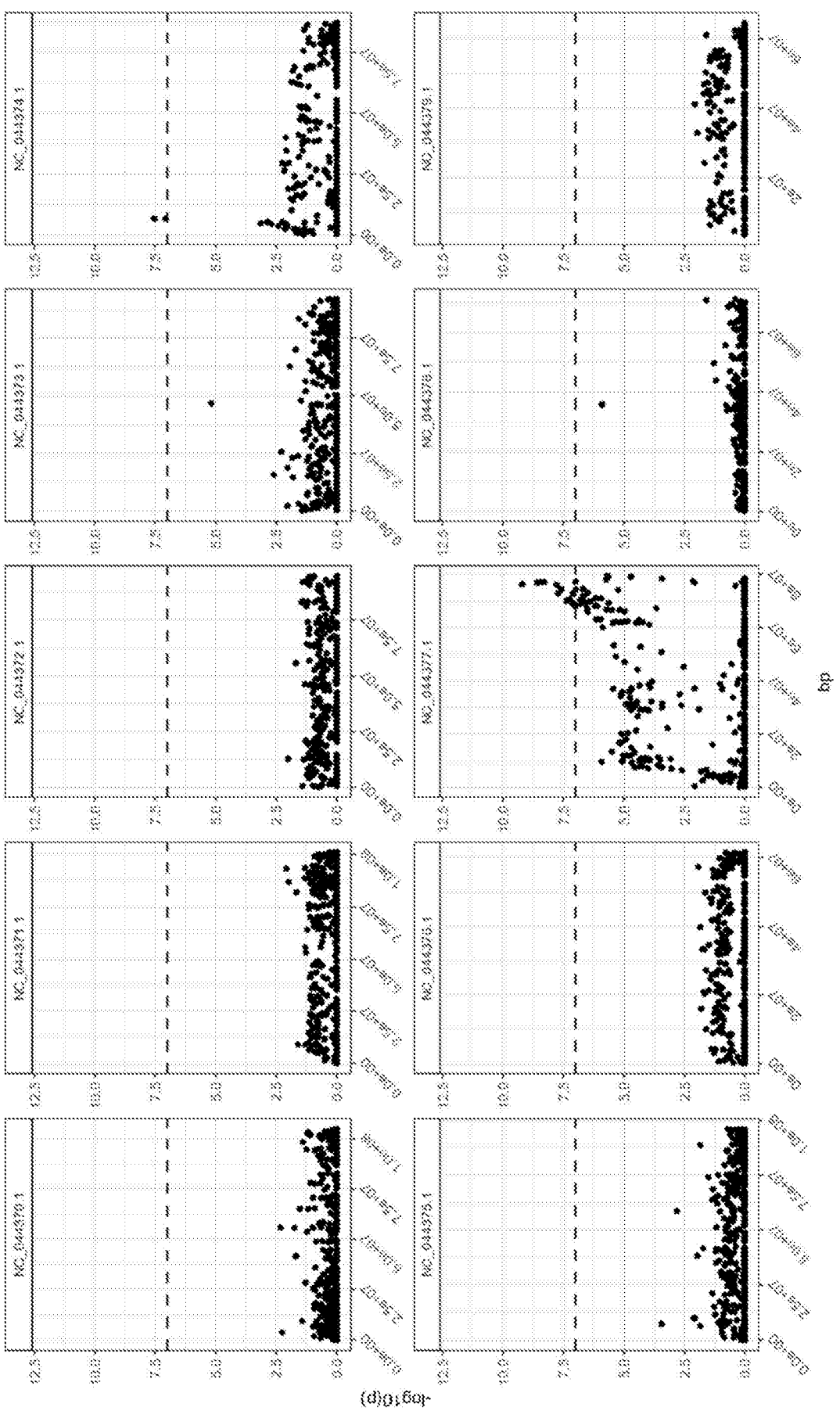
NC_044370.1
NC_044371.1
NC_044372.1
NC_044373.1
NC_044374.1
NC_044375.1
NC_044376.1
NC_044377.1
-log10(p)
bp

Figure 2

QUANTITATIVE TRAIT LOCI ASSOCIATED WITH PURPLE COLOR IN CANNABIS

BACKGROUND OF THE INVENTION

The present invention describes methods of identifying a Cannabis sativa plant comprising quantitative trait loci (QTLs) associated with purple color, and to Cannabis sativa plants comprising the QTLs. The invention also relates to plants with increased levels of purple color identified by the methods. The invention further relates to marker assisted selection and marker assisted breeding methods for obtaining plants having purple color, as well as to methods of producing Cannabis sativa plants with the absence of purple color and/or varying degrees of purple color and plants produced by these methods.

Modern Cannabis is derived from the cross hybridization of three biotypes; Cannabis sativa L. ssp. indica, Cannabis sativa L. ssp. sativa, and Cannabis sativa L. ssp. ruderalis. Cannabis was divergently bred into two distinct, albeit tentative types, called Hemp and HRT (high-resin-type) Cannabis, respectively, which are typically used for different purposes. Hemp is primarily used for industrial purposes, for example in feed, food, seed, fiber, and oil production. Conversely, HRT cannabis is largely cultivated and bred for high concentrations of the pharmacological constituents, cannabinoids, derived from resin in the trichomes. Biomass, including the leaf and stem, of cannabis can also be an important source of cannabinoids.

Cannabis is the only species in the plant kingdom to produce phytocannabinoids. Phytocannabinoids are a class of terpenoid acting as antagonists and agonists of mammalian endocannabinoid receptors. The pharmacological action is derived from this ability of phytocannabinoids to disrupt and mimic endocannabinoids. Due to its psychoactive properties, one cannabinoid, delta-9-tetrahydrocannabinol (THC), the decarboxylation product of the plant-produced delta-9-tetrahydrocannabinolic acid (THCA), has received much attention in illegal or unregulated breeding programs, with modern HRT varieties having THC concentrations of 0.5% to 30%.

Cannabis can display a multitude of colors in its leaves, stem and inflorescence. Purple color displayed by some cannabis strains is an important characteristic for visual appeal in markets for HRT Cannabis. Purple Haze, for example, is named and marketed, in part, for the purple color of its inflorescence. Purple color of flowers is also an undesirable trait in some cases, some consumers prefer HRT Cannabis flowers that are light or dark green that show no purple. This makes flower color an important trait for HRT cannabis breeders, producers, and consumers. Selection of cannabis with or without purple color can be challenging as breeders may have to wait for the emergence of the purple color, especially in flowers, toward the end of a plant's life cycle. The purple color in cannabis plants is most likely the product of anthocyanin accumulation.

Anthocyanins are water-soluble flavonoids. This class of small molecules absorb specific wavelengths of the electromagnetic spectrum depending on their chemical structure. The

absorbance of blue-green wavelengths of light by anthocyanins in plants can result in the appearance of purple color. Anthocyanin accumulates in the vacuole of epidermal cells conferring a range of colors, dark blue, purple, and reds, to plants. These colors can serve to attract pollinators and animal herbivores for seed dispersal. Anthocyanins may play important roles in plant stress mitigation to cold and drought, for example, by dampening the effect of reactive oxygen species. This suggests that purple color in cannabis plants may be an important trait for stress tolerance in HRT and Hemp cannabis.

The biosynthesis of anthocyanins has been well characterized in several plant species, though not in Cannabis. Anthocyanins are formed, like other flavonoids, from the coupling of three molecules of malonyl-CoA with 4-coumaryl CoA by Chalcone synthase to form naringenin chalcone. The isomerization of naringen chalcone is then catalyzed by chalcone isomerase (CHI) to naringenin. Naringenin is then oxidized by successive enzymes flavanone hydroxylase, flavonoid 3'-hydroxylase, and flavonoid 3',5'-hydroxylase. The products of these oxidations are then converted to colorless leucoanthocyanidins by dihydroflavonol 4-reductase (DFR) and subsequently to colored anthocyanidins by anthocyanidin synthase (ANS). Sugar molecules are then coupled to the unstable anthocyanidins by various members of the glycosyltransferase enzyme family, resulting in stable anthocyanins.

Anthocyanin biosynthesis can be induced by developmental cues in response to abiotic and biotic stress. MYB transcription factors, R2R3-MYBs and R3-MYBs, have been demonstrated to play roles in the regulation of anthocyanin biosynthesis, and in secondary metabolism in general, in many agronomically important plant species. MYB transcription factors can act as positive regulators of anthocyanin production, such as MYB10 that can regulate skin color of apple varieties by activating the expression of genes that encode proteins for anthocyanin biosynthesis. MYB transcription factors also act as negative regulators of anthocyanin biosynthesis. For example, the R2R2-Myb of *Brassica rapa*, BrMYB4 inhibits anthocyanin accumulation by repressing the expression of cinnamate 4-hydroxylase, required for the biosynthesis of 4-coumaryl CoA.

The genetic basis for the accumulation or absence of purple color in cannabis is not known. While anthocyanin accumulation is a likely cause of the presence of purple color, the mechanisms underlying its regulation are unclear. Though MYB transcription factors have been shown to play a role in the regulation of anthocyanin accumulation in other plants species, the size of this family of transcription factors and the diversity of the activities of its members make it impossible to infer the role of MYB transcription factors in Cannabis. In many plant species, fruit or flower color can be affected by unwanted excessive browning in tissue rich in anthocyanins, caused by polyphenol oxidase catalysing the degradation of anthocyanins to brown break down products. Understanding the genetic basis of purple color in cannabis can benefit the cannabis industry through elimination or inclusion of this trait to meet consumer preference. Regulation of this trait may also be important for developing climate resistant HRT and Hemp type cannabis

varieties. The identification of molecular markers for this trait can facilitate acceleration of breeding times for varieties selecting for multiple traits. The present invention relates to markers and the identity of putative genes for the control of purple color accumulation in cannabis.

SUMMARY OF THE INVENTION

The present invention relates to a method for identifying a Cannabis sativa plant comprising in its genome a purple color QTL, the alleles of which are either associated with the absence or the presence of a purple color trait in the plant. The invention further relates to methods of producing a Cannabis sativa plant comprising in its genome the purple color QTL. In addition, the present invention relates to Cannabis sativa plants identified or produced according to the methods disclosed and to Cannabis sativa plants containing a QTL associated with the presence or absence of purple color. Also provided are quantitative trait loci that control a purple color trait in Cannabis sativa, wherein the quantitative trait locus is defined by single nucleotide polymorphisms defined herein or genetic markers linked to the QTL, as well as putative genes that control a purple color trait in a Cannabis sativa plant.

According to a first aspect of the present invention there is provided for a method for identifying a Cannabis sativa plant comprising in its genome a genomic region including a purple color QTL, the method comprising the steps of: (i) genotyping at least one plant with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4; and (ii) identifying one or more plants containing the purple color QTL. In particular, the polymorphism may be selected from the group consisting of “common_4519”, “common_4525”, and “common_4500”, as defined in Table 4.

In a first embodiment of the method for identifying a Cannabis sativa plant comprising a purple color QTL, the genotyping may performed by PCR-based detection, including using molecular markers, sequencing of PCR products containing the one or more polymorphisms, targeted resequencing, whole genome sequencing, or restriction-based methods, for detecting the one or more polymorphisms. While many suitable genotyping methods are known to those of skill in the art, in one embodiment, the genotyping may be performed using sequencing primers or similar molecular markers, wherein the molecular markers may be selected from the primer pairs as defined in Table 5 herein, which have been developed by the inventors of the present invention for detecting the polymorphisms provided in Tables 1 to 4 herein.

According to a second embodiment of the method for identifying a Cannabis sativa plant comprising a purple color QTL, the molecular markers may be designed for detecting polymorphisms at regular intervals within the purple color QTL such that recombination can be excluded.

In a third embodiment of the method for identifying a Cannabis sativa plant comprising a purple color QTL, the molecular markers may be designed for detecting polymorphisms at regular intervals within the purple color QTL such that recombination can be quantified to estimate linkage

disequilibrium between a particular polymorphism and a purple color phenotype, or the absence thereof. For example, molecular markers may be for detecting polymorphisms such that recombination events can be detected to a resolution of 10'000 or 100'000 or 500'000 base pairs within the QTL.

According to a second aspect of the present invention, there is provided for a method of producing a Cannabis sativa plant having a genomic region including a purple color QTL in its genome, the method comprising the steps of: (i) providing a donor parent plant having in its genome a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4; (ii) crossing the donor parent plant having the purple color QTL with at least one recipient parent plant that does not have the purple color QTL to obtain a progeny population of cannabis plants; (iii) screening the progeny population of cannabis plants for the presence of the purple color QTL; and (iv) selecting one or more progeny plants having the purple color QTL, wherein the plant displays the purple color trait. In particular, the polymorphism characterizing the purple QTL may be selected from the group consisting of "common_4519", "common_4525", and "common_4500", as defined in Table 4.

In a first embodiment of the method of producing a Cannabis sativa plant having a purple color QTL, the method may further comprise the step of: (v) crossing the one or more progeny plants with the donor recipient plant; or (vi) selfing the one or more progeny plants.

According to a second embodiment of the method of producing a Cannabis sativa plant having a purple color QTL, the screening step may comprise genotyping at least one plant from the progeny population with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4.

In a third embodiment of the method of producing a Cannabis sativa plant having a purple color QTL, the method may comprise a step of genotyping the donor parent plant with respect to the purple color QTL prior to providing said donor plant, by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4.

According to an alternative aspect of the present invention there is provided for a method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the method comprising the steps of: (i) providing a donor parent plant having in its genome a QTL associated with an absence of purple color characterized by one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4; (ii) crossing the donor parent plant having the QTL associated with the absence of purple color with at least one recipient parent plant that has a purple color QTL to obtain a progeny population of cannabis plants; (iii) screening the progeny population of cannabis plants for the presence of the QTL associated with the absence of purple color; and (iv) selecting one or more progeny plants having the QTL associated with the absence of purple color, wherein the plant does not display the purple color trait.

In a first embodiment of the method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the method may further comprise: (v) crossing the one or more progeny plants with the donor recipient plant; or (vi) selfing the one or more progeny plants.

According to a second embodiment of the method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the step of screening may comprise genotyping at least one plant from the progeny population with respect to the QTL associated with the absence of purple color by detecting one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4.

In a further embodiment of the method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the method may further comprise a step of genotyping the donor parent plant with respect to the purple color QTL by detecting one or more polymorphisms associated with the presence or absence of purple color as defined in any one of Tables 1 to 4. In particular, the plant may be screened for a polymorphism selected from the group consisting of “common_4519”, “common_4525”, and “common_4500”, as defined in Table 4.

In another embodiment of both the method of producing a Cannabis sativa plant having a purple color QTL and the method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the genotyping may be performed by PCR-based detection using molecular markers, sequencing of PCR products containing the one or more polymorphisms, targeted resequencing, whole genome sequencing, or restriction-based methods, for detecting the one or more polymorphisms.

In some embodiments, the molecular markers may be for detecting polymorphisms at regular intervals within the QTL such that recombination can be excluded or such that recombination can be quantified to estimate linkage disequilibrium between a particular polymorphism and a purple color phenotype or absence of purple color phenotype. For example, molecular markers may be for detecting polymorphisms such that recombination events can be detected to a resolution of 10’000 or 100’000 or 500’000 base pairs within the QTL. In an alternative embodiment, genome sequencing, or marker-based PCR and resequencing of the QTL may be used for detecting a plurality of polymorphisms defined in any one of Tables 1 to 4. In some embodiments, the molecular markers may be selected from the primer pairs provided in Table 5. Further, in some embodiments, the progeny population of cannabis plants contains a minimum of 100, or 500, or 1000, or 10000 plants.

In a further aspect of the present invention, there is provided for a method of producing a Cannabis sativa plant comprising a purple color trait, the method comprising introducing a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4 into a Cannabis sativa plant, wherein said purple flower QTL is associated with the purple color trait.

In one embodiment of the method of producing a Cannabis sativa plant comprising a purple color trait, introducing the purple color QTL may comprise crossing a donor parent plant in which the purple color QTL is present, with a recipient parent plant in which the purple color QTL is not present.

In an alternative embodiment of the method of producing a Cannabis sativa plant comprising a purple color trait, introducing the purple color QTL may comprise genetically modifying the Cannabis sativa plant. Several methods of genetic modification are known to those of skill in the art, including targeted mutagenesis, genome editing, and gene transfer. For example, one or more of the polymorphisms as defined in any one of Tables 1 to 4 herein may be introduced into a plant by mutagenesis and/or gene editing, in particular the methods of genetically modifying a plant may be selected from the group consisting of CRISPR-Cas9 targeted gene editing, heterologous gene expression using various expression cassettes; TILLING, and non-targeted chemical mutagenesis using e.g. EMS. Alternatively, a cannabis sativa plant may be transformed with the purple color QTL or a part thereof, via any of the transformation methods known in the art.

In an alternative aspect of the invention there is provided for a method of producing a Cannabis sativa plant that does not display a purple color trait, the method comprising introducing a QTL characterized by one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4 into a Cannabis sativa plant, wherein said QTL is associated with the absence of purple color in the plant.

In one embodiment of the method of producing a Cannabis sativa plant that does not display a purple color trait, introducing the QTL may comprise crossing a donor parent plant in which the QTL associated with the absence of purple color is present, with a recipient parent plant in which the QTL is not present.

In an alternative embodiment of the method of producing a Cannabis sativa plant that does not display a purple color trait, introducing the QTL associated with the absence of purple color may comprise genetically modifying the Cannabis sativa plant. Several methods of genetic modification are known to those of skill in the art, including targeted mutagenesis, genome editing, and gene transfer. For example, one or more of the polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4 herein may be introduced into a plant by mutagenesis and/or gene editing, in particular the methods of genetically modifying a plant may be selected from the group consisting of CRISPR-Cas9 targeted gene editing, heterologous gene expression using various expression cassettes; TILLING, and non-targeted chemical mutagenesis using e.g. EMS. Alternatively, a cannabis sativa plant may be transformed with the QTL associated with the absence of purple color or a part thereof, via any of the transformation methods known in the art.

According to a further aspect of the present invention there is provided for a Cannabis sativa plant identified according to any method of identifying a Cannabis plant described herein,

or produced according to any method of producing a Cannabis plant described herein, provided that the plant is not exclusively obtained by means of an essentially biological process.

In yet a further aspect of the present invention there is provided for a Cannabis sativa plant comprising a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4, provided that the plant is not exclusively obtained by means of an essentially biological process.

In an alternative aspect of the invention there is provided for a Cannabis sativa plant comprising a QTL associated with the absence of purple color characterized by one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4, provided that the plant is not exclusively obtained by means of an essentially biological process.

According to another aspect of the present invention there is provided for a quantitative trait locus that controls a purple color trait in Cannabis sativa, wherein the quantitative trait locus is defined by a single nucleotide polymorphism at position 80922439 of NC_044373.1 or a genetic marker linked to the QTL; or wherein the quantitative trait locus is defined by a single nucleotide polymorphism at position 6600328 of NC_044374 or a genetic marker linked to the QTL; or wherein the quantitative trait locus has a sequence that corresponds to nucleotides 68717484 to 77040783 of NC_044377.1 and is defined by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4 or a genetic marker linked to the QTL. The invention further include a genomic region defined by markers linked to the QTLs defined herein.

In yet a further aspect of the present invention there is provided for an isolated gene that controls a purple color trait in a Cannabis sativa plant, wherein the gene is selected from the group consisting of the genes as defined in Table 6 with reference to the CS10 genome.

In one embodiment, the isolated gene has the gene identity number LOC115695758 and encodes a putative MYB Transcription factor, as defined in Table 6.

In another embodiment, the isolated gene has the gene identity number LOC115695872 or LOC115695871 and encodes an anthocyanidin 3-O-glucosyltransferase 2, as defined in Table 6.

BRIEF DESCRIPTION OF THE FIGURES

Non-limiting embodiments of the invention will now be described by way of example only and with reference to the following figures:

Figure 1: GWA of Purple Color in Cannabis in a F2 Population.

Figure 2: GWA for Validation of Purple Color in Cannabis in a F2 Population.

SEQUENCES

The nucleic acid and amino acid sequences listed herein and in any accompanying sequence listing are shown using standard letter abbreviations for nucleotide bases, and the

standard one or three letter abbreviations for amino acids. It will be understood by those of skill in the art that only one strand of each nucleic acid sequence is shown, but that the complementary strand is included by any reference to the displayed strand.

DETAILED DESCRIPTION OF THE INVENTION

The present invention will now be described more fully hereinafter with reference to the accompanying drawings, in which some, but not all embodiments of the invention are shown.

The invention as described should not be limited to the specific embodiments disclosed and modifications and other embodiments are intended to be included within the scope of the invention. Although specific terms are employed herein, they are used in a generic and descriptive sense only and not for purposes of limitation.

As used throughout this specification and in the claims, which follow, the singular forms "a", "an" and "the" include the plural form, unless the context clearly indicates otherwise.

The terminology and phraseology used herein is for the purpose of description and should not be regarded as limiting. The use of the terms "comprising", "containing", "having" and "including" and variations thereof used herein, are meant to encompass the items listed thereafter and equivalents thereof as well as additional items.

Methods are provided herein for identifying and obtaining plants having a purple color trait prior to the plant displaying the color phenotypically, using a molecular marker detection technique. The inventors of the present invention have further produced purple colored cannabis plants by crossing plants displaying purple color to cannabis plants lacking purple color. Also demonstrated herein, the inventors were able to use genome wide association (GWA) to identify multiple QTLs linked to purple color. This finding provides for the improvement of methods for producing plants displaying differing degrees of purple color and plants that do not display purple color.

A total of three QTLs for purple color were identified in the mixed populations tested and the two F2 populations tested.

Tables 1 to 4 herein provides several single nucleotide polymorphisms (SNPs) which define the QTLs associated with the purple color. In some embodiments one or more of the identified SNPs can be used to incorporate the purple color trait from a donor plant, containing one or more of the QTLs associated with the trait, into a recipient plant. For example, the incorporation of the purple color phenotype may be performed by crossing a donor parent plant to a recipient parent plant to produce plants containing a haploid genome from both parents. Recombination of these genomes provides F1 progeny where each haploid complement of chromosomes, of the diploid genome, is comprised of genetic material from both parents.

In some embodiments, methods of identifying one or more QTLs that are characterized by a haplotype comprising of a series of polymorphisms in linkage disequilibrium are provided. The QTLs each display limited frequency of recombination within the QTLs. Preferably the

polymorphisms are selected from any one of Tables 1 to 4 herein, representing the purple color QTLs. Molecular markers may be designed for use in detecting the presence of the polymorphisms and thus the QTLs. Further, the identified QTL polymorphisms and the associated molecular markers may be used in a cannabis breeding program to predict the purple color trait of plants in a breeding population and can be used to produce cannabis plants that either display the purple color trait, or do not display the purple color trait, compared to a control population.

As used herein, reference to a "purple color" plant or a variety with a "purple color" trait refers to a plant or a variety that has the appearance of purple color at the time of harvest. In particular, a plant of purple color may accumulate a higher level of anthocyanin or anthocyanin-related compounds compared to a plant that does not have purple color at the time of harvest.

The time of harvest is defined with respect to the maturity of the flower, where approximately greater than 50% of the pistils have turned brown in appearance. The time of harvest can also be determined by initiation of flowering for hemp-type cannabis or by other agronomic criteria common in the art.

As used herein a "quantitative trait locus" or "QTL" is a polymorphic genetic locus with at least two alleles that differentially affect the expression of a continuously varying phenotypic trait when present in a plant or organism which is characterised by a series of polymorphisms in linkage disequilibrium with each other.

As used herein, the term "purple color QTL" or "purple color quantitative trait locus" refers to a quantitative trait locus comprising part, or all, of the QTLs characterized by the polymorphisms described in any one of Tables 1 to 4.

As used herein, "haplotypes" refer to patterns or clusters of alleles or single nucleotide polymorphisms that are in linkage disequilibrium and therefore inherited together from a single parent. The term "linkage disequilibrium" refers to a non-random segregation of genetic loci or markers. Markers or genetic loci that show linkage disequilibrium are considered linked.

As used herein, the term "purple color haplotype" refers to the subset of the polymorphisms contained within the purple color QTLs which exist on a single haploid genome complement of the diploid genome, and which are in linkage disequilibrium with the purple color trait.

As used herein, the term "donor parent plant" refers to a plant that is either homozygous or heterozygous for the purple color haplotype or which contains one or more of the purple color QTLs. Alternatively, the donor parent plant may be one that is not heterozygous homozygous for the purple color QTL, or the purple color haplotype, where the absence of the purple color trait is desirable.

As used herein, the term "recipient parent plant" refers to a plant that is not heterozygous or homozygous for the purple color QTL, or the purple color haplotype. Alternatively, the recipient parent plant may be one that is either homozygous or heterozygous for the purple color haplotype

or which contains one or more of the purple color QTLs, where the absence of the purple color trait is desirable.

The term "crossed" or "cross" means the fusion of gametes via pollination to produce progeny (e.g., cells, seeds or plants). The term encompasses both sexual crosses (the pollination of one plant by another) and selfing (self-pollination, e.g., when the pollen and ovule are from the same, or genetically identical plant). The term "crossing" refers to the act of fusing gametes via pollination to produce progeny.

The term "purple color allele" refers to the haplotype allele within a particular QTL that confers, or contributes to, the purple color phenotype, or alternatively, is an allele that allows the identification of plants with the purple color phenotype that can be included in a breeding program ("marker assisted breeding" or "marker assisted selection").

The term "GWAS" or "Genome wide association study" or "GWA" or "Genome wide association" as used herein refers to an observational study of a genome-wide set of genetic variants or polymorphisms in different individual plants to determine if any variant or polymorphism is associated with a trait, specifically the purple color trait.

As used herein a "polymorphism" is a particular type of variance that includes both natural and/or induced multiple or single nucleotide changes, short insertions, or deletions in a target nucleic acid sequence at a particular locus as compared to a related nucleic acid sequence. These variations include, but are not limited to, single nucleotide polymorphisms (SNPs), indel/s, genomic rearrangements, gene duplications, as well as genome insertions and deletions.

As used herein, the term "LOD score" or "logarithm (base 10) of odds" refers to a statistical estimate used in linkage analysis, wherein the score compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. The LOD score is a statistical estimate of whether two genetic loci are physically near enough to each other (or "linked") on a particular chromosome that they are likely to be inherited together. A LOD score of 3 or higher is generally understood to mean that two genes are located close to each other on the chromosome. In terms of significance, a LOD score of 3 means the odds are 1,000:1 that the two genes are linked and therefore inherited together.

As used herein, the term "quantile-quantile" or "Q-Q" refers to a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

As used herein, a "causal gene" is the specific gene having a genetic variant (the "causal variant") which is responsible for the association signal at a locus and has a direct biological effect on the purple color phenotype. In the context of association studies, the genetic variants which are responsible for the association signal at a locus are referred to as the "causal variants". Causal

variants may comprise one or more “causal polymorphisms” that have a biological effect on the phenotype.

The term “nucleic acid” encompasses both ribonucleotides (RNA) and deoxyribonucleotides (DNA), including cDNA, genomic DNA, isolated DNA and synthetic DNA. The nucleic acid may be double-stranded or single-stranded. Where the nucleic acid is single-stranded, the nucleic acid may be the sense strand or the antisense strand. A “nucleic acid molecule” or “polynucleotide” refers to any chain of two or more covalently bonded nucleotides, including naturally occurring or non-naturally occurring nucleotides, or nucleotide analogs or derivatives. By “RNA” is meant a sequence of two or more covalently bonded, naturally occurring or modified ribonucleotides. The term “DNA” refers to a sequence of two or more covalently bonded, naturally occurring or modified deoxyribonucleotides. By “cDNA” is meant a complementary or copy DNA produced from an RNA template by the action of RNA-dependent DNA polymerase (reverse transcriptase).

The term “isolated”, as used herein means having been removed from its natural environment. Specifically, the nucleic acid or gene(s) identified herein may be isolated nucleic acids or gene(s), which have been removed from plant material where they naturally occur.

The term “purified”, relates to the isolation of a molecule or compound in a form that is substantially free of contamination or contaminants. Contaminants are normally associated with the molecule or compound in a natural environment, purified thus means having an increase in purity as a result of being separated from the other components of an original composition. The term "purified nucleic acid" describes a nucleic acid sequence that has been separated from other compounds including, but not limited to polypeptides, lipids and carbohydrates which it is ordinarily associated with in its natural state.

The term “complementary” refers to two nucleic acid molecules, e.g., DNA or RNA, which are capable of forming Watson-Crick base pairs to produce a region of double-strandedness between the two nucleic acid molecules. It will be appreciated by those of skill in the art that each nucleotide in a nucleic acid molecule need not form a matched Watson-Crick base pair with a nucleotide in an opposing complementary strand to form a duplex. One nucleic acid molecule is thus “complementary” to a second nucleic acid molecule if it hybridizes, under conditions of high stringency, with the second nucleic acid molecule. A nucleic acid molecule according to the invention includes both complementary molecules.

As used herein a “substantially identical” or “substantially homologous” sequence is a nucleotide sequence that differs from a reference sequence only by one or more conservative substitutions, or by one or more non-conservative substitutions, deletions, or insertions located at positions of the sequence that do not destroy or substantially reduce the antigenicity of the expressed fusion protein or of the polypeptide encoded by the nucleic acid molecule. Alignment for purposes of determining percent sequence identity can be achieved in various ways that are within the knowledge of those with skill in the art. These include using, for instance, computer

software such as ALIGN, Megalign (DNASTAR), CLUSTALW or BLAST software. Those skilled in the art can readily determine appropriate parameters for measuring alignment, including any algorithms needed to achieve maximal alignment over the full length of the sequences being compared. In one embodiment of the invention there is provided for a polynucleotide sequence that has at least about 80% sequence identity, at least about 90% sequence identity, or even greater sequence identity, such as about 95%, about 96%, about 97%, about 98% or about 99% sequence identity to the sequences described herein.

Alternatively, or additionally, two nucleic acid sequences may be “substantially identical” or “substantially homologous” if they hybridize under high stringency conditions. The “stringency" of a hybridisation reaction is readily determinable by one of ordinary skill in the art, and generally is an empirical calculation which depends upon probe length, washing temperature, and salt concentration. In general, longer probes required higher temperatures for proper annealing, while shorter probes require lower temperatures. Hybridisation generally depends on the ability of denatured DNA to re-anneal when complementary strands are present in an environment below their melting temperature. A typical example of such “stringent” hybridisation conditions would be hybridisation carried out for 18 hours at 65 °C with gentle shaking, a first wash for 12 min at 65 °C in Wash Buffer A (0.5% SDS; 2XSSC), and a second wash for 10 min at 65 °C in Wash Buffer B (0.1% SDS; 0.5% SSC).

Methods of identifying a QTL or haplotype responsible for the purple color phenotype and molecular markers therefor

In some embodiments, methods are provided for identifying a QTL or haplotype responsible for purple color and for selecting plants with the purple color trait. In some embodiments, the methods may comprise the steps of:

a. Identifying a plant that displays the purple color phenotype within a breeding program.

b. Establishing a population by crossing the identified plant to itself (selfing) or a recipient parent plant.

c. Genotyping the resultant F1, or subsequent populations, for example by sequencing methods.

d. Performing association studies, including phenotyping and linkage analysis, to discover QTLs and/or polymorphisms contained within the QTL.

e. Optionally, identifying cannabis paralogs of previously characterized genes that may be involved in the purple color phenotype.

f. Developing molecular markers that detect one or more polymorphisms linked to QTLs, alleles within these QTLs, or existing or induced polymorphisms.

g. Validating the molecular markers by determining the linkage disequilibrium between the marker and the purple flower trait.

Trait development and introgression

In some embodiments, methods are provided for marker assisted breeding (MAB) or marker assisted selection (MAS) of plants having a purple color QTL or trait. The methods may comprise the steps of:

a. Identifying a plant that displays the purple color trait or phenotype or contain a purple color QTL as defined herein.

b. Establishing a population by crossing the identified plant to itself (selfing) or another recipient parent plant.

c. Genotyping and phenotyping the resultant F1, or subsequent, populations, for example by sequencing methods.

d. Performing association studies, inputting phenotype and genotype information to identify genomic regions enriched with polymorphisms associated with the purple color trait, to discover QTLs and/or polymorphisms contained within the QTL.

e. Optionally, identifying cannabis paralogs of previously characterized genes that may be involved in the purple color phenotype.

f. Developing molecular markers that detect one or more polymorphisms linked to QTLs, alleles within these QTLs, or existing or induced polymorphisms.

g. Using the molecular markers when introgressing the QTLs or polymorphisms into new or existing cannabis varieties to select plants containing the purple color haplotype or the purple color trait, or plants where the purple color haplotype or the purple color trait is absent.

QTLs and Marker Assisted Breeding

In some embodiments, during the breeding process, selection of plants displaying the purple color trait may be based on molecular markers designed to detect polymorphisms linked to genomic regions that control the trait of interest by either an identified or an unidentified mechanism. Previously identified genetic mechanisms may, for example, have a direct or pleiotropic effect on purple color in a plant. Examples include genes selected from: MYB transcription factors, such as R2R3-MYBs, R3-MYBs, MYB10, R2R2-MYBs, including BrMYB4. In some embodiments, QTLs containing such elements are identified using association studies. Knowledge of the mode-of-action is not required for the functional use of these genomic regions in a breeding program. Identification of regions controlling unidentified mechanisms may be useful in obtaining plants with the purple color phenotype, based on identification of polymorphisms that are either linked to, or found within QTLs that are associated with the purple color phenotype using AS.

Construction of breeding populations

Breeding populations are the offspring of sexual reproduction events between two or more parents. The parent plants (F0) are crossed to create an F1 population each containing a chromosomal complement of each parent. In a subsequent cross (F2), recombination has occurred and allows for mostly independent segregation of traits in the offspring and importantly the reconstitution of recessive phenotypes that existed in only one of the parental lines.

According to some embodiments, QTLs that lead to the purple color phenotype are identified within synthetic populations of plants capable of revealing dominant, recessive, or complex traits. In one embodiment of the invention, a genetically diverse population of cannabis varieties, that are used to produce the synthetic population are integrate them into a breeding program by unnatural processes. In some embodiments, these processes result in changes in the genomes of the plants. The changes may include, but are not limited to, mutations and rearrangements in the genomic sequences, duplication of the entire genome (polyploidy), or activation of movement of transposable elements which may inactivate, activate or attenuate the activity of genes or genomic elements. According to one embodiment of the invention, the following methods are employed to integrate the plants into a breeding program include some or all of the following:

a. Growing plants in rich media or soils under artificial lighting;

b. Cloning of plants, often through a multitude of sub-cloning cycles;

c. Introduction of plants into in vitro, sterile growth environments, and subsequent removal to standard growth conditions;

d. Exposure to mutagens such as EMS, colchicine, silver nitrate, ethidium bromide, dinitroanalines, high concentrations mono or poly-chromatic light sources;

e. Growing plants under highly stressful conditions which include restricted space, drought, pathogen, atypical temperatures, and nutrient stresses.

Purple color trait association studies and QTL identification

In some embodiments, the synthetic populations created are either the offspring of the sexual reproduction or clones of plants in the breeding program such that genetic material of individuals in the synthetic populations is derived from one, or two, or more plants from the breeding program.

In one embodiment, plants identified within the synthetic population as having a trait of interest, such as the purple color trait, may be used to create a structured population for the identification of the genetic locus responsible for the trait. The structured population may be created by crossing one (selfing) or more plants and recovering the seeds from those plants.

Plants in the structured population may be fully genotyped using genome sequencing to identify genetic markers for use in the association study (AS) database. Association mapping is a powerful technique used to detect quantitative trait loci (QTLs) specifically based on the statistical correlation between the phenotype and the genotype. In this case the trait is the purple

color phenotype. In a population generated by crossing, the amount of linkage disequilibrium (LD) is reduced between genetic marker and the QTL as a function of genetic distance in cannabis varieties with similar genome structures. Simple association mapping is performed by biparental crosses of two closely related lines where one line has a phenotype of interest and the other does not. In some embodiments, advanced population structures may be used, including nested association mapping (NAM) populations or multi-parent advanced generation inter-cross (MAGIC) populations, however it will be appreciated that other population structures can also be effectively used. Biparental, NAM, or MAGIC structured populations can be generated and offspring, at F1 or later generations, may be maintained by clonal propagation for a desired length of time. In some embodiments, QTLs may be identified using the high-density genetic marker database created by genotyping the founder lines and structured population lines. This marker database may be coupled with an extensive phenotypic trait characterization dataset, including, for example, the purple color phenotype of the plants. Using the association studies described herein, together with accurate phenotyping, this method is able to identify genomic regions, QTLs and even specific genes or polymorphisms responsible for the purple color phenotype that are directly introduced into recipient lines. Polygenic phenotypes may also be identified using the methods described herein.

In one embodiment, the structured population is grown to the flowering stage. To characterize the phenotypes of the lines they are clonally reproduced so the phenotypic data can be collected in feasible replicates.

Molecular Markers to detect polymorphisms

As used herein, the term “marker” or “genetic marker” refers to any sequence comprising a particular polymorphism or haplotype described herein that is capable of detection. For example, a marker may be a binding site for a primer or set of primers that is designed for use in a PCR-based method to amplify and thus detect a polymorphism or haplotype. Alternatively, the marker may introduce a restriction enzyme recognition site, or result in the removal of a restriction enzyme recognition site. Plants can be screened for a particular trait based on the detection of one or more markers confirming the presence of the polymorphism. Marker detection systems that may be used in accordance with the present invention include, but are not limited to polymerase chain reaction (PCR) followed by sequencing, Kompetitive allele specific PCR (KASP), restriction fragment length polymorphisms (RFLPs) analysis, amplified fragment length polymorphisms (AFLPs), cleaved amplified polymorphic sequences (CAPS), or any other markers known in the art.

In some embodiments “molecular markers” refers to any marker detection system and may be PCR primers, such as those described in the examples below. For example, PCR primers may be designed that consist of a reverse primer and two forward primers that are homologous to the part of the genome that contains a polymorphism but differ in the 3’ nucleotide such that

the one primer will preferentially bind to sequences containing the polymorphism and the other will bind to sequences lacking it. The three primers are used in single PCR reactions where each reaction contains DNA from a cannabis plant as a template. Fluorophores linked to the forward primers provide, after thermocycling, a different relative fluorescent signal for homozygous and heterozygous alleles containing the polymorphism and for those lacking the polymorphism, respectively.

In some embodiments, allele-specific primers may each harbor a unique tail sequence that corresponds with a universal FRET (fluorescence resonant energy transfer) cassette. For example, the primer specific to the SNP may be labelled with a FAM and the other specific primer with a HEX dye. During the PCR thermal cycling performed with these primers, the allele-specific primer binds to the genomic DNA template and elongates, so attaching the tail sequence to the newly synthesized strand. The complement of the allele-specific tail sequence is then generated during subsequent rounds of PCR, enabling the FRET cassette to bind to the DNA. Alleles are discriminated through the competitive binding of the two allele-specific forward primers. At the end of the PCR reaction a fluorescent plate is read using standard tools which may include RT-PCR devices with the capacity to detect florescent signals and is evaluated with commercial software.

If the genotype at a given polymorphism site is homozygous, one of the two possible fluorescent signals will be generated. If the genotype is heterozygous, a mixed fluorescent signal will be generated. By way of example, genomic DNA extracted from cannabis leaf tissue at seedling stage can be used as a template for PCR amplifications with reaction mixtures containing the three primers. Final fluorescent signals can be detected by a thermocycler and analyzed using standard software for this purpose, which discriminates between individuals that are heterozygotes or homozygotes for either allele.

In some embodiments, molecular markers to one, two or more of the SNPs in the haplotype can be used to identify the presence of the QTL and by association, the purple color phenotype.

Further, the QTL may include a number of individual polymorphisms in linkage disequilibrium, which constitute a haplotype and which, with high frequency, can be inherited from a donor parent plant as a unit. Therefore, in some embodiments, molecular markers can be utilized which have been designed to identify numerous polymorphisms which are in linkage disequilibrium with other polymorphisms, any of which can be used to effectively predict the purple color phenotype of the offspring.

According to some embodiments, any polymorphism in linkage disequilibrium with one or more of the purple color QTLs can be used to determine the presence or absence of the haplotype in a breeding population of plants, as long as the polymorphism is unique to the purple color trait in the donor parent plant when compared to the recipient parent plant.

In some embodiments of the invention, the donor parent plant is a plant that has been genetically modified to include a purple color QTL defined by a polymorphism, for example any or all of the polymorphisms of any one of Tables 1 to 4. In an alternative embodiment, where the desired trait is the absence of the purple color trait, the donor parent plant may be a plant that has been genetically modified to exclude a purple color QTL defined by a polymorphism, for example any or all of the polymorphisms of any one of Tables 1 to 4.

In some embodiments, donor parent plants, as described above, are used as one of two parents to create breeding populations (F1) through sexual reproduction. Methods for reproduction that are known in the art may be used. The donor parent plant provides the trait of interest to the breeding population. The trait is made to segregate through the population (F2) through at least one additional crossing event of the offspring of the initial cross. This additional crossing event can be either a selfing of one of the offspring or a cross between two individuals, provided that each plant used in the F1 cross contains at least one copy of a purple color QTL allele or purple color haplotype, where the presence of the purple color trait is desirable.

In some embodiments, the presence or absence of the purple color allele or purple color haplotype in plants to be used in the F1 cross is determined using the described molecular markers. In some embodiments, the resulting F2 progeny is/are screened for any of the purple color polymorphisms described herein.

The plants at any generation can be produced by asexual means like cutting and cloning, or any method that yields a genetically identical offspring.

Production of purple color Cannabis sativa plants or Cannabis sativa plants lacking the purple color trait

In some embodiments, a Cannabis sativa plant that does not have the purple color trait may be converted into a purple color plant according to the methods of the present invention by providing a breeding population where the donor parent plant contains a purple color QTL associated with the purple color trait and recipient parent plant does not display the purple color phenotype.

In alternative embodiments, a Cannabis sativa plant that has the purple color QTL associated with the purple color trait may be converted into a plant lacking the purple color phenotype according to the methods of the present invention by providing a breeding population where the donor parent plant does not contain a purple color QTL associated with the purple color trait and recipient parent plant has a QTL associated with the purple color trait.

In some embodiments the purple color phenotype may be introduced into a recipient parent plant by crossing it with a donor parent plant comprising the purple color phenotype. In some embodiments the donor parent plant comprises a purple color phenotype and a contiguous genomic sequence characterized by one or more of the polymorphisms of any one of Tables 1 to 4.

In an alternative embodiment the purple color phenotype may be removed from a recipient parent plant by crossing it with a donor parent plant lacking the purple color phenotype. In some embodiments the donor parent plant lacks a purple color phenotype and a contiguous genomic sequence characterized by one or more of the polymorphisms of any one of Tables 1 to 4.

In some embodiments, the donor parent plant is any cannabis variety that is cross fertile with the recipient parent plant.

In some embodiments, MAS or MAB may be used in a method of backcrossing plants carrying or lacking the purple color trait to a recipient parent plant. For example, an F1 plant from a breeding population can be crossed again to the recipient parent plant. In some embodiments, this method is repeated.

In some embodiments, the resulting plant population is then screened for the purple color trait using MAS with molecular markers to identify progeny plants that contain or lack one or more purple color polymorphisms, such as those described in any one of Tables 1 to 4, indicating the presence or absence of an allele of a QTL associated with the purple color phenotype. In another embodiment, the population of cannabis plants may be screened by any analytical methods known in the art to identify plants with desired characteristics.

Methods to genetically engineer plants to achieve the presence or absence of purple color using mutagenesis or gene editing techniques

Identifying QTLs, and individual polymorphisms, that correlate with a trait when measured in an F1, F2, or similar, breeding population indicates the presence of one or more causative polymorphisms in close proximity the polymorphism detected by the molecular marker. In some embodiments, the polymorphisms associated with the purple color trait is introduced into, or removed from, a plant by other means so that a trait, such as the purple color trait, can be introduced into plants that would not otherwise contain associated causative polymorphisms or removed from plants that would otherwise contain associated causative polymorphisms.

The entire QTLs or parts thereof which confer the purple color trait described herein may be introduced into, or removed from, the genome of a cannabis plant to obtain plants with or without a purple color phenotype through a process of genetic modification known in the art, for example, but not limited to, heterologous gene expression using various expression cassettes.

The trait described herein may be introduced into, or removed from, the genome of a cannabis plant to obtain plants that include or exclude the causative polymorphisms and the potential to display a purple color phenotype through processes of genetic modification known in the art, for example, but not limited to, CRISPR-Cas9 targeted gene editing, TILLING, non-targeted chemical mutagenesis using e.g. EMS.

Plants may be screened with molecular markers as described herein to identify transgenic individuals with or without a purple color QTL or polymorphism(s), following the genetic modification.

In some embodiments, cannabis plants comprising or lacking one or more of the polymorphisms of any one of Tables 1 to 4 associated with the purple color QTLs are provided. In some embodiments the purple color QTL, or one or more polymorphisms associated therewith are introduced into, or removed from, the plants. For example, by genetic engineering. In some embodiments the one or more polymorphisms are introduced into, or removed from, the plants by breeding, such as by MAS or MAB, for example as described herein.

Accordingly, in a further embodiment, *Cannabis sativa* plants comprising or lacking a purple color QTL described herein, or one or more polymorphisms associated therewith, are provided, with the proviso that the plant is not exclusively obtained by means of an essentially biological process.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLE 1

Genome-wide association studies (GWAS) of purple flower color in Cannabis

During outdoor field trials in 2020 it was observed that several populations of cannabis plants were comprised of individuals with varying degrees of purple-colored flowers. To identify molecular markers for the appearance of purple color in Cannabis the study was initially focused on the apical inflorescence in a diverse population comprising 3220 individuals.

Trimmed and dried apical inflorescence of Cannabis sativa genotypes were photographed and visually assessed for the presence of purple areas.

Individual plants whose apical inflorescence showed at least some purple areas were coded as 1, those only showing green areas were coded as 0.

DNA was extracted from about 70 mg of leaf discs from all the plants evaluated using an adapted kit with “sbeadex” magnetic beads by LGC Genomics, which was automated on a KingFisher Flex with 96 Deep-Well Head by Thermo Fisher Scientific.

The extracted DNA served as a template for the subsequent library preparation for sequencing. The library pools were prepared according to the manufacturer’s instructions (AgriSeq™ HTS Library Kit—96 sample procedure from Thermo Fisher Scientific). Targeted sequencing of a custom SNP marker panel based on the Cannabis Sativa CS10 reference genome was carried out on the Ion Torrent system by Thermo Fisher Scientific. The primers for the SNPs identified are provided in Table 5. The library pool was loaded onto Ion 550 chips with Ion Chef and sequenced with Ion GeneStudio S5 Plus according to the manufacturer’s instructions (Ion 550™ Kit from Thermo Fisher Scientific).

From a population of 3220 individuals, a genome-wide association study (GWAS) was performed to detect significant associations between genotypic information derived from targeted resequencing of the custom SNP marker panel described above and the appearance of purple

color in the apical inflorescence. Flowers were coded as 1 for those showing at least some purple and 0 for those only showing green areas.

The genotypic matrix was filtered for SNPs having more than 30% missing values within the population and a minor allele frequency lower than 5%. This resulted in 2699 SNP markers after filtering. The GWAS was performed using GAPIT version 3 (J. Wang & Zhang, 2021) with five statistical models: General Linear Model (GLM), Mixed Linear Model (MLM), FarmCPU and Blink (model=c("GLM", "MLM", "FarmCPU", "Blink"). A quantile-quantile plot (QQ plot) was used to evaluate the statistical models. The Blink model performed the best by our evaluation and was used for the analysis. SNPs surpassing a LOD ($-\log_{10}$(p-value)) value of 5 were considered to have a significant association with trait variation.

SNPs showing a significant association with purple color in flower, with an LOD value greater than 5, were found on chromosome NC_044371.1, NC_044373.1, NC_044377.1 with reference to the Cannabis Sativa CS10 genome and are listed in Table 1. The homozygous allele of the SNPs in Table 1 that can distinguish the presence of purple flower are listed along with their position and reference sequence. Interestingly the heterozygous state is also indicative of purple flower color, however less so than the homozygous state of the allele for purple flower color, indicating this is a dominant trait.

Table 1: SNPs associated with the purple color trait in flower field trial. The presence of the purple color trait is predicted by the occurrence of the indicative allele (marked with *). The positions of the SNPs are provided with reference to the CS10 reference genome as described herein. “Homo_1” denotes the average phenotypic value associated with homozygous allele 1 based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant, “Homo_2” denotes the average phenotypic value associated with homozygous allele 2 based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant and “Hetero” denotes the average phenotypic value associated with heterozygous based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant. BP refers to the nucleotide position of the SNP.

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_4485	NC_044377.1	72948533	20,0423	A	G *	0,1	0,21	0,17	CGGCTACGCTTTCGCCGGGGATAGCTCTCTTCCCGCGCCGGTCATATT TTCCGGCGTTCCTCCGGCGACAACCGCTACTGCCACCGCTTGGTCACC TTCCTTATCGTCTGCTCTCTACAAAGTCGATGGGTGGGGCGCACCTTAC TTCGCCGTCAACTCTTCCGGCAACGTCGCCGTTCGCCCTCACGGCGCT GGAACTTT[G/A]GCGCACCAGGAGATTGATTTGCTGAAAATTGTAAAGAA GGTTTCGGATCCGAAATCTAAGGGCGGGTTAGGTTTGCCCCTTCCGCT CGTTATTCGGCTTCCTGATGTGCTTAAGAACCGCCTTGAGTCTCTCCAG GCGGCGTTCGATTTCGCAATCCAGTCGCAGGACTATGAAAACCATTACC AGGGTGTTTACCCTGTG **(SEQ ID NO:1)**
common_2262	NC_044373.1	80922439	8,277951	A*	G	0,2	0,12	0,18	TCAAGAAATAAGAATTAACTAAATATTGCCACTTAACCTAGTAAAATTAA GAGCAGTTTCACGTGTTAAAATTAAATAAAATAATGAACTCAACAGAAAC TTAAACCACAATACCTCCAACTTGTTGGGGGGTATTGGTTTTACATTCTC TACTTTCAATGTACAACAACCAACAGTATCTGCTTCATCATCATCCTGCA A[G/A]GCATGAGATCATAATGTCAGTGATGCCTTCAGTTAACAGCTAGGG GAACTTAAAGATGATGTAGAGAAACATGTTGAGCAACTTAAATGATGCA ATGTAGCAGATTCGAGCAACAAATATAAATCAAACACCTTTCATATTTCT TTTTCCTTAACAAGGGCACAAATAAACAATGACAAAAATGTTAAACTGAA GAGCTGT **(SEQ ID NO:2)**
common_2032	NC_044373.1	44537208	7,285258	A*	C	0,19	0,14	0,18	CATGCAAATACATACATATACATATGTGATTTTTTTTTTATGTATACTATAT ATATAGGTCATGGGGTGATGTGAATGAGTGCTTGGGAAGCAAGAGATC AAGTATCATAAATAATGGAGGGGAGTTAGGCCATGAATATCGTGATCAT CATCATCATCAGCATCGGCATATTGGGGTTTCAACGATGAAGATGAAGA AGAT[A/C]AAAGGGAGAAGGAAAGTGAGAGAACCAAGGTTTTGCTTTAA GACCATGAGCGAGGTCGATGTGCTTGATGATGGTTACAAGTGGAGAAA GTACGGACAGAAAGTGGTCAAGAACACACAGCATCCCAGGTACCTAATT AATATCCATTTATTCATTAATTTATAATATTAACAACTCTAACAATGCCATT AATATTAAAGC **(SEQ ID NO:3)**
common_5220	NC_044379.1	34679389	6,59189	A*	G	0,2	0,13	0,14	GATTGCACTGTGAAAGGGATTCGAAGTAACGGCCGGATCCTCATGACG GTCTTGATCAATATGAGCAATTAATGTTCTAGGGTTTAGGCCTCTTGATT GATGGTGGTGATGGGCAGCCGCCATTGAAACGGAATTACTAGGGTTTT CTAATTCCCTGGTCGTAGTAGTGTTTGATGTAGTGGTTGCCATGAGAAA ATGGGA[A/G]GTAGGCTTTCGTTCTTGGATTATTTGGGTGGTGCTAGTAG TAGTACTACTACTCATTATAAGCTCTGGCTTATTATTCATGTCACAAAAT GAGTAGTGTGGTTTTGATGAGAAGAAGAAGCTGGGTGATGTTGGGTTAA TGAGTTCCCAGTTCCAAGGAGCAGGTAATAAACCTAGAAAGTCGTCCTC GGACTTTGACCCA **(SEQ ID NO:4)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_816	NC_044371.1	45301500	6,588524	A	G*	0,15	0,28	0,21	TCCAATCTTTGGGAGCAACACTGTAAGATGTAACAGTGACACC TTCACTATTAGTGACTTCAAAGGATAGGGGTTGATTTTTTAAAT CAGCATTTATGTGCCAGTTTTGGCCCCAGTTCCTTCCCATTGA AAGCCAACCAGTTCTTGAACCCTTAATTTTCACAGCAGATACG TCTCCAGCACCGGCAACATTGCTGATTA[A/G]CACCGAAATGA AGATACTTGAACCGTCAATTGTGAATCGTATTCCTCCTTCCTTT CTGCACTTTATTCTGTTACAGAAGTCAAATTAACTAGTTGTCAA CATGCCACATTTTTTCCTAATAGATTTAATCACACCGAACATAG TTTTCGTCATCAATATGCCACATTTACAACAACATTGTCATCTC ATCTTTTTCACTAT **(SEQ ID NO:5)**
common_4499	NC_044377.1	74270487	6,51065	A	G*	0,04	0,21	0,13	CGAAAAGCCTATCTGAAATCTTAGTTCTACAGATAGGACACTC TGAACAAGCTAATGAACAAGATTTACATACTGCAAAACAAAAT GATAAATATGATTAGTTTTAATTTAATGTTAAAAGTGAAATCAA GCATGTAGGATTCATAGGAATGATGATGACTGCTTACAACAAA AATGGCGACACGGCAAAAGAATTGCAGC[G/A]GTCGGGGATT CAAAACATACTTTACACATGTGAGAATTGGCATCTCCATTCCC CATTTGTTTCAGCTCCTTTTCCTTCATCTCTTGCATTCGAGCCT GAAAACAACATACATCTTGAGTTTGAGATACCTTAACACTTTTT CGAGAACTTTTTAGTTTCTTGATCAACAAATGCATTTTCCTTGT GGATTTAAAATTATGC **(SEQ ID NO:6)**
common_4448	NC_044377.1	68957824	6,199807	A	G*	0,16	0,18	0,15	TTGATCAGCGAAGAAAGGCCAACCAATAATTGGCACTCCAGC GCTCAAACTCTCCAGTGTCGAGTTCCAACCACAATGCGATAAA AACCCTCCTATTGCAGGGTGGTTCAAAACCTCTTCTTGTGGAC ACCAACTACAAATCACACATCTTTCTCTTGTCTCTTCCACAAAC TCAGTTGGAATAATTCCCGAGTTTCCATC[G/A]ATGATGTCGG GTCTTATAATCCAAACAAAGGGTTTCCCACTGTTAGCCAAACC CCAAGCGAACTCAACAAGTTCGTCAGGTGTCATAGCCGTGAT GCTGCCGAAATTAACATAAATCACCGAATTGGCCTCCCTAGAA TTCAACCATTGCAAGCATTCAAGCTCTTCTTTCCATAGATTAGA TCCAATGGATGACAAACTT **(SEQ ID NO:7)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_4452	NC_044377.1	69326994	5,601726	A*	G	0,26	0,11	0,2	GAGTGCCGTATATTTGTATTTAAACATTAGTCAACCAA TATGATCAAATGTATATATACGGTTACATATTACGCAT ATATATGAATCAAAGTTATATTACTTTCTCAATATGATC AAAGTTGTGATTTTGTTGGTGCTAGCCACACTCATTAA TCAAGTATGGAGTAGTACTACAAGTAATAATTATTGCA TAGAGAAGGA[G/A]AGACAAGCTCTTCTCAACTTGAAG AAAGGCTTTGTCGATGATGGCAATCGTCTATCCTCAT GGACAAGTAGTAGCCGTGATTGTTGTGCATGGAGAG GTATCAGGTGCGATAACTCAAAAACTCATCGTCATATT ATCGCTCTTGATCTTAAATCTGATGACAACAATCATAA TTATTTGGGTGGTGAAATTGGTCCTTCT **(SEQ ID NO:8)**
GBScompat_rare_2	NC_044370.1	519152	5,110202	A*	G	0,25	0,16	0,24	GTCGCAAATGGAAATTTACGCCCGCGATGTACTCGAA TTAAACCCATTAACCCCATTTCTCAGGTACTCCACAAA ATCATCATATTACTTTTTCTTTCTAATTTCACATTTTTTT GAATTTGTTTTTGGGTTTTGGTGGAAATAGGTGAAAG GATTGCCATTTAATCGGTATTCATGGCTAACAACCCA CAATGCGTTTGC[G/A]AAGCTGGGACAGAAATCGCAG ACGGGAACACCGATTGTGTCTTCCATGAATCAACAGG ACTCCATTACTAGCCAGCTCAATGTAAGTTTTTTTTTTT TTCTTTTTAGTTAGTGAAATTTATGTTGTTTGTTTCTCG GGAAAGTTTTGCGGTAAAATTAAGGGGAAAATATGAT CAATGACTGGACTTTACAATAACTAAAA **(SEQ ID NO:9)**

EXAMPLE 2

Genome-wide association studies (GWAS) of whole plant purple color in Cannabis

It was observed that the purple color in cannabis is not restricted to the flowers alone. It can be found in leaves, stem, and other components of the shoot system of cannabis. The inventors thus sought to identify additional SNP markers associated with whole plant purpleness and to understand if the markers found associated with purple color in flowers were also relevant to the presence of purple color in the whole plant. They assessed purple color visually of the whole plant from a mixed population, that is a subset of the population used in Example 1, consisting of 2274 individuals.

At the time of harvest, plants were photographed, and genotypes were visually assessed for the presence of purple in the whole plant, the areas on leaf, stem, and flower. Plants showing at least some purple areas were coded as 1, those only showing green areas were coded as 0.

DNA was extracted from about 70 mg of leaf discs from all the plants evaluated using an adapted kit with “sbeadex” magnetic beads by LGC Genomics, which was automated on a KingFisher Flex with 96 Deep-Well Head by Thermo Fisher Scientific.

The extracted DNA served as a template for the subsequent library preparation for sequencing. The library pools were prepared according to the manufacturer’s instructions (AgriSeq™ HTS Library Kit—96 sample procedure from Thermo Fisher Scientific). Targeted sequencing of a custom SNP marker panel based on the Cannabis Sativa CS10 reference genome was carried out on the Ion Torrent system by Thermo Fisher Scientific. The primers for the SNPs identified are provided in Table 5. The library pool was loaded onto Ion 550 chips with Ion Chef and sequenced with Ion GeneStudio S5 Plus according to the manufacturer’s instructions (Ion 550™ Kit from Thermo Fisher Scientific).

From a population of 2274 individuals, a genome-wide association study (GWAS) was performed to detect significant associations between genotypic information derived from targeted resequencing of the custom SNP marker panel described above and the appearance of purple color in the whole plant. Plants were coded as 1 for those showing at least some purple and 0 for those only showing green areas.

The genotypic matrix was filtered for SNPs having more than 30% missing values within the population and a minor allele frequency lower than 5 %. This resulted in 2350 SNP markers after filtering. The GWAS was performed using GAPIT version 3 (J. Wang & Zhang, 2021) with five statistical models: General Linear Model (GLM), Mixed Linear Model (MLM), FarmCPU and Blink (model=c("GLM", "MLM", "FarmCPU", "Blink"). A quantile-quantile plot (QQ plot) was used to evaluate the statistical models. The Blink model performed the best by our evaluation and was used for the analysis. SNPs surpassing a LOD ($-\log_{10}$(p-value)) value of 5 were considered to have a significant association with trait variation.

The inventors identified SNPs significantly associate with purple color in the whole plant on chromosome NC_044372.1, NC_044377.1, NC_044378.1, listed in Table 2. They identified two SNP markers that were found in both experiments “common_4485” and “common_4448”, as well as 10 additional SNP markers. The new insight indicated that the same QTL on chromosome NC_044377.1 was associated with purple color in both the flower and the whole plant.

Table 2: SNPs associated with the purple color trait in a whole plant field trial. The presence of the purple color trait is predicted by the occurrence of the indicative allele (marked with *). The positions of the SNPs are provided with reference to the CS10 reference genome as described herein. Homo_1” denotes the average phenotypic value associated with homozygous allele 1 based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant, “Homo_2” denotes the average phenotypic value associated with homozygous allele 2 based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant and “Hetero” denotes the average phenotypic value associated with heterozygous based on scoring for purple color from 0 to 1, where 1 indicated a purple plant and 0 indicated a green plant. BP refers to the nucleotide position of the SNP.

SNP	Chromosome	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4451	NC_044377.1	69163028	17,16585	A	C*	0,07	0,25	0,15	AAATGGCACCGCATCCGAAACAACAAATATCCCAATCAAACGA AAGAACTCACTTATTGCCTTATGACACCGTCTTGCTTCATCCCC ACTCTGATCATCGCTACCCGATGCTCCGAAATAACGCTTTCCA GCCACCATTCTCACCACCATGTTAAGAGTTAAATCTTCTAACCA ACGACTCAACTCAACTACAACTTTACA[A/C]GACTTGTAGAGCT CTCTAATCCCTACTTCAACCTCTGAAATCCTCACTTGCTTCAAC ATCTCTAGACGGCGGTTAGAGAGGAGTTCTAACGTGGCGATCT TCCTCATTTCGCGCCAAAAAGGGCTATAAGGTGCGAAACCAAA GACTGCGTAGTTGTAGCCCATGTGCTTGGCTGCCACGGTTGTA GGGCGCGAGGCCAGC **(SEQ ID NO:10)**
common_4502	NC_044377.1	74475495	11,38578	C*	G	0,21	0,15	0,19	GATAATTTGATCTTACCTTGTTGTGTACCATAAGTAATCAGTGG AGTCTCTGGTGACATTTAACTGCTCTAAGAGACCAACTGCTGT GATTTTTGTATCTTCATCCACTGAAGACATGTCTTCACTAAATGT CTCCCATGAGTACAATGCACCTTTCAAGGGCGACATTTGCGTT CGAGTCATTTGTACTCTCACCTGTTA[G/C]GAAGTGACAGATGC CCGTTATAGACAATGAGTGAAGTAGCCAATGCATCTTACATATA GTACACGATCAACCCTGTGATAAAGCGGGCTAGCCCTGATATG GTATTATGTTCTTATATCAAGCTCAATAATTAATCGTTTTCTTAT GGCCCATTTAGTCTCAGGGTCAGTCCTGTTATAGTATTATGTTC TTACACATTAT **(SEQ ID NO:11)**
GBScompat_rare_165	NC_044377.1	73090006	8,571359	A	G*	0,17	0,25	0,35	CTTCAGCCCAAACCTTTTGAGCAACCTCAAAAGTCCTATCAATA CCATGCTCACCATCAAGTAAGATCTCTGGCTCAACAATTGGGA CCAAACCATTGTCCTAAACATAACAAAACATAATCAATTTTCACA AACACAAGTACAAAAATCAAGATTAGCTTAGTCTTGTCTGATTA GTTATCTTACTTGGGAGATAGCAGC[A/G]TAGCGGGCTAGACC CCAAGCGGCTTCCTTCACAGCCAGAGCTGATGGGCCATTGGG AATGCTCACAACAGTACGCCTGTTTAAATAATGTAATTTTGTGT CAAATACCGATTAAAGAGCATTGCTATAAGGCATTATAGTGTCC AACATCACTATTAGGCACTACGGGTGCTTTTTAACATTTCTCAA CCAGACAAAAGTA **(SEQ ID NO:12)**

SNP	Chromosome	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4500	NC_044377.1	74383124	7,631759	A	C*	0,11	0,25	0,19	TCGACTTCAATGAGAAATCTCAACCCCTGTGGTAAGAATTTTG TGTACTTTTTTAAAAATTGAATTTTTAAAATTTCCCAGTTCGCTG AATATTTGTATTAACTGTGTCTAATTTTTCATACTAGACATTGAA AGAATGGTGTCTTTGAAGGGAATGGTTATCCGTTGCAGTTCAA TAATTCCGGAAATCAGAGAAGCAAT[C/A]TTTAGATGTCTTGTG TGCGGTTACTACTCTGAACCTGTTGCTGTAGAGAAAGGTAATT AGTTAGGATACCATTCGCATGGGCTAGTTTTTTCTTATCTATAT CACAATGTCATTCTAAATTATTTTCCCTTTTCAGGACGGATAAC TGAACCAACAAGATGCTTGAAGGAAGAATGCCAAGCAAGAAA CTCCATGACACTT **(SEQ ID NO:13)**
common_4054	NC_044377.1	7972754	7,433539	A*	G	0,17	0,15	0,23	TTAACTTAATGATCATATAGATAGTAATAAACTATTAATTAATTA ATTTTGCTGGGGATGAGAATGGTGGCCAGGTAGCTTTTCCTT GATCTTTTCCATAACAGTTTTCTTCTCGTGAAGAGGAGTAGGA TCATGAGTAGTAGCAGTTGGTTCGTCCTTGTGTTTTCCAGTAA GTTTCTCTTTTATTTTTGTTCCTAAACC[C/T]TTCTTCTTCCTTCT CCCTCCAGCTCCATCATCCTCTGACTGCATCCACACCCATTTT ATAATTATTATATATATTAATCATTCATTTTTAATATATATAATTA CATTATATTATTATTTTAATATTTTAAAATGTAAATAAAAAATTTA GTGATTCAAATTATAATTTTTTATATTATATGTTGATAGCAATTT TGTTATT **(SEQ ID NO:14)**
common_4448	NC_044377.1	68957824	6,644629	A	G*	0,15	0,2	0,18	TTGATCAGCGAAGAAAGGCCAACCAATAATTGGCACTCCAGC GCTCAAACTCTCCAGTGTCGAGTTCCAACCACAATGCGATAAA AACCCTCCTATTGCAGGGTGGTTCAAAACCTCTTCTTGTGGAC ACCAACTACAAATCACACATCTTTCTCTTGTCTCTTCCACAAAC TCAGTTGGAATAATTCCCGAGTTTCCATC[G/A]ATGATGTCGG GTCTTATAATCCAAACAAAGGGTTTCCCACTGTTAGCCAAACC CCAAGCGAACTCAACAAGTTCGTCAGGTGTCATAGCCGTGAT GCTGCCGAAATTAACATAAATCACCGAATTGGCCTCCCTAGAA TTCAACCATTGCAAGCATTCAAGCTCTTCTTTCCATAGATTAGA TCCAATGGATGACAAACTT **(SEQ ID NO:7)**

SNP	Chromosome	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4519	NC_044377.1	76201790	5,94937	A*	G	0,26	0,09	0,19	CGATCACTTCGTAGATGCATCCTCCCACAAGGTAGCACAATTGT AGAAAGTGCTAAATCATGCTTTATTCCATTTGTTCTTTTTGTCTTC TCTTTTTGCTTAATCGAACGATGTTGTGAACTTGTAGGGTTGTCA AATTCGAAGGGAGAGCGCACATGGAGAGAGCGTTTGTTGCAACA ATGTGAGGGCCCTGTTTGATGA[A/G]CTCCCAACTCCACACCTAA TTGTGGAGATCACACCATTCCCTGAAGGGCCTCTCACTGAAAAA GATTACACCAAAGCTGAGAAATTGGAGAGGGTACTTAGAACTGG CCCGAACGTTTGATTCTTCTCTCGAGTTAAATCATCGCTGTCTCT CGTTAGAACTACAGCTTAATTGTATGTATGTTTTGAGCCTTGTAC ATAT **(SEQ ID NO:15)**
common_4599	NC_044378.1	3495196	5,705698	A	T*	0,17	0,21	0,14	CCATATAATGCAAATTCTCTAATATCAATTACAAAATCAAATATAA GACACAGATGCCTAAATGATGCCATGTAATTTCGATCACAGCATT GAATTTTTCTTCAAGATTGAAGAGTTAAGAAGTAAAGAACTTTAC CATAGATGTGTAGGAGACATTGTAACGAAAAAGAGGCGCTTCTT GCGAGGATCAACTTTAGTAGC[A/T]ACCCAATCTGCCCAAGCTTC CATAGCCAACTCCATGGCACCAACACCATCTAACTCTTCACAAAT TCCACTTTCATCAGAACTCCATCTGCATAATAAATTGGTTGCATA GTAAGTACAAAGTGCTCAACACATCATATTCTGAAACTTAACTCA AAATGTTACTATGATTCATTTTACTTACAACAGCTTAACTGGACCT **(SEQ ID NO:16)**
common_1535	NC_044372.1	63488008	5,52871	A	G*	0,144	0,19	0,23	GAAGTATGGGAAAGAAAGAGCTTGAGTCAAGAAGCTTTGATTGA GGCATTCAAAATGGAAGTGAATCAAGTAAAAATGGAGTTAAGAA GTGCAAATGAGTTGGTGAGCCAGTTGATGGATGATGTTGAAATA CTTACTTATGATATACTAAAGGCGAAAACTGAGATTAATGAAATG AAGAGGAAAGAAACAGAAGCTCAA[A/G]TTGAGATAGCACTAATG AAAACTGAGCTTCAGAAAGAGAGAAGAGATTGCAATGATGGTCA CATAACGGCTTCACTCAAGGAATACGAGTCCTTGGTCAAAAAGG GTGATGATCAAATTTGGGCGCCAACTTCTGATAACAAGCATGAG CTGGAAACTTTGAGGAAGGAGTTGGATGCTGCATTGGCTAAAGT TGCTGAAT **(SEQ ID NO:17)**

SNP	Chromosome	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4485	NC_044377.1	72948533	5,369889	A	G*	0,11	0,24	0,18	CGGCTACGCTTTCGCCGGGGATAGCTCTCTTCCCGCGCCGGT CATATTTTCCGGCGTTCCTCCGGCGACAACCGCTACTGCCACC GCTTGGTCACCTTCCTTATCGTCTGCTCTCTACAAAGTCGATG GGTGGGGCGCACCTTACTTCGCCGTCAACTCTTCCGGCAACG TCGCCGTTCGCCCTCACGGCGCTGGAACTTT[G/A]GCGCACCA GGAGATTGATTTGCTGAAAATTGTAAAGAAGGTTTCGGATCCG AAATCTAAGGGCGGGTTAGGTTTGCCCCTTCCGCTCGTTATTC GGCTTCCTGATGTGCTTAAGAACCGCCTTGAGTCTCTCCAGGC GGCGTTCGATTTCGCAATCCAGTCGCAGGACTATGAAAACCAT TACCAGGGTGTTTACCCTGTG **(SEQ ID NO:1)**
common_4446	NC_044377.1	68780117	5,323884	A	T*	0,026	0,23	0,09	TTTGTTAATTGTTTCATTTTTCTGAAATGCAAATGATGATATTTTT ATGAAGGTTGGAAGAATGGCGGGTCAGTTTGCCAAGCCCAGA TCAGATTCATTTGAGGATAGGGATGGAGTGAAGCTTCCTAGTT ACAGAGGTGACAATGTGAATGGTGATGCATTTGATGAAAAATC GAGAATTCCAGACCCTAATCGAATGAT[T/A]AGAGCCTATACAC AATCTGCTGCAACTTTGAATCTTTTGAGGGCATTTGCTACTGGA GGTTATGCTGCTATGCAGAGAGTGACCCACTGGAATCTAGATT TCACTGATCACAGTGAGCAGGGGGATAGGTAAACTTTTATTGT TCTTTTCTTCTTACTTGAAATTTTGAATGTTTATTTTCCATAATGA ATAGGATTGAAG **(SEQ ID NO:18)**
common_4514	NC_044377.1	75675432	5,184746	A*	G	0,31	0,11	0,18	CACTTTAAGTTATAAATTACGTTGTAACTAAAAGTAAAAATCTTT GTAGTGTAAATTTATATATATTTACCTCGGAGACCATGTCATTG AAAACTCTTCCCATTATTTCCATCTTATGTTTAGGATCATTACTC ATAACACTTCCCATCATGGTTAGTACCATACTCCCACCACTTAC TATTTCCTCTGAACGACATCTTA[A/G]AAAGCTTGTGAAATCCTC TTGAAATTGATTCAAGTAAGCTTTCACAACTGAATTAGGGCTTC TTTTTGTAATATAAATGTTGCCTTTGTTAAGTGCTTCTCCTGTTT CCCCAACCAAACCACTTGGAACCTTCAAAACAACATTAATTCAC ATAATTAATAAACTAAAACCTTATAAAAAGAAAGGGTAAATTTCA ATTTT **(SEQ ID NO:19)**

EXAMPLE 3

Genome-wide association studies (GWA) of purple color in an F_2 Population in Cannabis

To confirm the ability to monitor the transmissibility of the purple color through monitoring SNP markers associated with this trait in the next generations, to identify additional SNP markers associated with purple color in cannabis, and to identify candidate genes that may be involved in the presence or absence of purple color, the inventors generated an F_2 population designated GID: 21 002 035 0000 from the selfing of a progeny from parents GID:20 000 104 0000 known to be stable for the appearance of purple color in the whole plant and GID:20 000 072 0000 known to rarely display purple color in the whole plant.

They assessed purple color visually of the whole plant from F_2 population GID: 21 002 035 0000 consisting of 137 individuals. At the time of harvest, genotypes were visually assessed for the presence of purple in the whole plant, the areas on leaf, stem, and flower.

Plants were assessed for purple color in the whole plant with a score from 1 to 9, where 1 indicates a completely green plant and 9 a completely purple plant. A total of 41 (28,87 % of total population) plants were scored less than 5 (predominantly green), while 101 (71,12 % of total population) plants were scored greater than or equal to 5 (more purple). This indicates a dominant allele controlling purple color in the whole plant and the flower and that the trait is transmissible.

DNA was extracted from about 70 mg of leaf discs from all the plants evaluated using an adapted “sbeadex kit” with magnetic beads by LGC Genomics, automated on a KingFisher Flex with 96 Deep-Well Head by Thermo Fisher Scientific.

The extracted DNA served as a template for the subsequent library preparation for sequencing. The library pools were prepared according to the manufacturer’s instructions (AgriSeq™ HTS Library Kit—96 sample procedure from Thermo Fisher Scientific). Targeted sequencing of a custom SNP marker panel based on the Cannabis Sativa CS10 reference genome was carried out on the Ion Torrent system by Thermo Fisher Scientific. The primers for the SNPs identified are provided in Table 5. The library pool was loaded onto Ion 550 chips with Ion Chef and sequenced with Ion GeneStudio S5 Plus according to the manufacturer’s instructions (Ion 550™ Kit from Thermo Fisher Scientific).

From a population of 137 individuals, a genome-wide association analysis (GWAS) was performed to detect significant associations between genotypic information derived from targeted resequencing of the custom SNP marker panel described above and the appearance of purple color in the whole plant. Plants were assigned a score from 1 to 9, where 1 indicates a completely green plant and 9 a completely purple plant.

The genotypic matrix was filtered for SNPs having more than 30% missing values within the population and a minor allele frequency lower than 5 %. This resulted in 4212 SNP markers after filtering. The GWAS was performed using GAPIT version 3 (J. Wang & Zhang, 2021) with five statistical models: General Linear Model (GLM), Mixed Linear Model (MLM), FarmCPU and

Blink (model=c("GLM", "MLM", "FarmCPU", "Blink"). A quantile-quantile plot (QQ plot) was used to evaluate the statistical models. The Blink model performed the best by our evaluation and was used for the analysis. SNPs surpassing a LOD ($-\log_{10}$(p-value)) value of 5 were considered to have a significant association with trait variation.

Here, SNPs significantly associated with purple color in the whole plant were found exclusively on chromosome NC_044377.1, listed in Table 3. The inventors show that the presence of the indicative homozygous allele is strongly associated with purple color in this segregating population. The SNPs identified in Table 3 are useful in predicting the presence or absence of purple color in the whole plant. The inventors show that the heterozygous state of the allele associates with purple color, though less so than the homozygous state. The homozygous state of the reference allele is clearly associated with plants that are not purple.

From SNP marker "common_4519" (Table 3), showing the highest association purple color in the F_2 population, a constant decrease in LOD values can be observed for neighboring increasingly distant markers, showing an erosion of linkage disequilibrium caused by recombination. This observation shows the marker panel used in this study can be used to monitor linkage decay across the genome and for determining the QTL with high confidence.

Table 3: SNPs associated with the purple color trait in a whole plant, F2 population 21 002 035 0000 on Chromosome NC_044377,1. The presence of the purple color trait is predicted by the occurrence of the indicative allele (marked with *). The positions of the SNPs are provided with reference to the CS10 reference genome as described herein. Homo_1" denotes the average phenotypic value associated with homozygous allele 1 based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant, "Homo_2" denotes the average phenotypic value associated with homozygous allele based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant and "Hetero" denotes the average phenotypic value associated with heterozygous based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant. BP refers to the nucleotide position of the SNP.

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4519	76201790	10,96014	A	G*	2,18	7,4	6,74	CGATCACTTCGTAGATGCATCCTCCCACAAGGTAGCACAATTGTAGAAAGTGCTAAA TCATGCTTTATTCCATTTGTTCTTTTTGTCTTCTCTTTTTGCTTAATCGAACGATGTTG TGAACTTGTAGGGTTGTCAAATTCGAAGGGAGAGCGCACATGGAGAGAGCGTTTGT TGCAACAATGTGAGGGCCCTGTTTGATGA[A/G]CTCCCAACTCCACACCTAATTGTG GAGATCACACCATTCCCTGAAGGGCCTCTCACTGAAAAAGATTACACCAAAGCTGAG AAATTGGAGAGGGTACTTAGAACTGGCCCGAACGTTTGATTCTTCTCTCGAGTTAAA TCATCGCTGTCTCTCGTTAGAACTACAGCTTAATTGTATGTATGTTTTGAGCCTTGTA CATAT **(SEQ ID NO:15)**
common_4525	76757669	10,61882	A	C*	2,38	7,76	6,72	ACAAGTCTTATCTAAACGAAAAGCTCCACAACTATGGAGATGAACACTACCAGAGAA GCAAAACATAACAGTTAGCTTCAGTTCATAATTTATTTACAGTCTATCATACACTGTTC TACAAGTGCTTGCTGAGGATGGATTCGACAACACCTTGTGCAATAGGATGATAACAA TCGCGAGCCTCTGCAAACACCCTTTTCG[C/A]GAAAATCTTCTCTTCTTCCTTCCCGT TCCCTTGAACGAGCGCAGTGTAGAGTGGACGAAGGTATTTCATCCTCCCAACTTCTT TGAGAGTTTTCTCCACCTCGCCGTAGTAGTCTCTGCACTTGGCGGTAATGGCTAGCT GCAGAAACCCCACCTTCACCTCGTAATCTTTTGATTCTGAGAGCCTGTAGCGCTGGT CCAA **(SEQ ID NO:20)**
common_4500	74383124	9,883966	A*	C	7,35	2,37	6,86	TCGACTTCAATGAGAAATCTCAACCCCTGTGGTAAGAATTTTGTGTACTTTTTTAAAA ATTGAATTTTTAAAATTTCCCAGTTCGCTGAATATTTGTATTAACTGTGTCTAATTTTTC ATACTAGACATTGAAAGAATGGTGTCTTTGAAGGGAATGGTTATCCGTTGCAGTTCA ATAATTCCGGAAATCAGAGAAGCAAT[C/A]TTTAGATGTCTTGTGTGCGGTTACTACT CTGAACCTGTTGCTGTAGAGAAAGGTAATTAGTTAGGATACCATTCGCATGGGCTAG TTTTTTCTTATCTATATCACAATGTCATTCTAAATTATTTTCCCTTTTCAGGACGGATAA CTGAACCAACAAGATGCTTGAAGGAAGAATGCCAAGCAAGAAACTCCATGACACTT **(SEQ ID NO:13)**
common_4487	73084792	9,515762	A*	G	7,41	2,75	6,64	CTGTCTCATATATTCAACCTCGTCACAAACATTTATTGTTTCGTGGGTCGTTTTCCAC TTCTTCATATCTACACAATAATCAATCGGGGCTTCCTTTGTTTTTTGTTATGTGAAAAT GGTTGCAGAAGTAGAAGAAGAATGAGAGAGTTAAGATTAGAAAGAAAGAAGAATGC GAACCCTTTGTGACGTTTGTGAGAAAGC[T/C]GCCGCGATCCTCTTCTGCGCCGCCG ACGAGGCTGCTCTATGCAGTTCCTGCGATGACAAGGTCTTTCACTAATTTAAGCTTTT AATAATTTTTAATTTAATTTACTAACTTCAAATCATTGGTATGAAGTTATTTCAATTCAA ATATTATTTATATGAATTGGGTTTTGTATTTTGTTATCTGCTGGAATTTTCATTGTTT **(SEQ ID NO:21)**

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
GBScompat_rare_165	73090006	9,271999	A	G *	2,55	7,32	6,76	CTTCAGCCCAAACCTTTTGAGCAACCTCAAAAGTCCTATCAATACCATGCTCAC CATCAAGTAAGATCTCTGGCTCAACAATTGGGACCAAACCATTGTCCTAAACAT AACAAAACATAATCAATTTTCACAAACACAAGTACAAAAATCAAGATTAGCTTAG TCTTGTCTGATTAGTTATCTTACTTGGGAGATAGCAGC[A/G]TAGCGGGCTAGA CCCCAAGCGGCTTCCTTCACAGCCAGAGCTGATGGGCCATTGGGAATGCTCA CAACAGTACGCCTGTTTAAATAATGTAATTTTGTGTCAAATACCGATTAAAGAGC ATTGCTATAAGGCATTATAGTGTCCAACATCACTATTAGGCACTACGGGTGCTT TTTAACATTTCTCAACCAGACAAAAGTA **(SEQ ID NO:12)**
GBScompat_rare_164	72805722	9,055252	A *	C	7,27	2,67	6,6	AGCAGAGAAAATTGAAAAAATGACGGAAAGGAAAAAGAAGAGGGAAGCAACAC CAGTATGCTTTGTGGCCACAGGGAAGTGAACAAGGTATAAATCCAGATAATCTA ACTGAAGCTTCTTCAGACTATTCTTACAGGCCTCAAGGACATGTCCATGGTCAG AATTCCAAAGCTGCAATCAAGACAGGTAAAGAAAAAGAAT[A/C]ATAGTCTGGCT GGACATACAGAATATACATATTATTCGATCAGTTCGAAAATAAACTAACCTTAGT GGTAACAAAGAGATCTTCTCTCTTAACAAGCCCTGTCTGAAATGCCTCAGAAAG TGCCTCCCCAACTTCTGTTTCATTCCTGTAATCAGCTGCAACCAACACCAGCTA GTCAAACATTACATTGTAGCAACTCAA **(SEQ ID NO:22)**
common_4513	75673037	8,673232	A	C *	2,15	7,42	6,63	ATAATGGGTTTTGATTTGTTCTGATACTCAGATTGAACCAAAGAGAAGGCCATC TCCGAGTTACCTTGAGAAAGTTCAGAGTGAAATCAGTGCCAACATGAGAGGAG TATTGGTGGATTGGTTGGTGGAAGTTGCAGAGGAGTACAAACTTGGTTCAGAG ACTCTTTACCTATCTGTTTCTTATATTGATCGATTCTTGTC[A/C]TTGAACACCAT TGCCAGGAATAAGCTTCAGTTATTGGGTGTTTCTTCTTTGCTCGTTGCCTCGTA AGATTCTAACCCTTTTGAACTAATGTTAATGAAGATGATATGTTGAATTTGATTT GTTTATTCATATAAAGTTTTGATTTTTATCTTTGGTTTCACTTGTTTAGAAAGTAT GAAGAGATTAATCCTCCTAATGTGG **(SEQ ID NO:23)**
common_4522	76434921	8,634652	A	G *	2,56	7,47	6,33	GACCTGGCGCCGGATCGATCCGGTTCCGAAATACGCGGACGGGTTGCCTATG TTCTGCCACGTGGCGGGTTGTGAAGGTAAGCTGGTTGTAATGGGCGGTTGGG ATCCGAAAAGCTACGGACCCGTTTCGGATGTTTTCGTTTTCGATTTCGCTAAGA ATCGGTGGAGCGAAGCCAAGGCCATGCCGGGTAAGAGGTCGTT[C/T]TTTGCG GTCGGGTCTTATTCGGGTCGGATCTATGTTGCGGGTGGGCACGATGAGAATAA GAACGCGTTGAGTTCTGCTTGGGTTTACGATGTGAGTCTGGACGAGTGGAGCG AGTTGGCTCAGATGAGTCAGGGCCGTGACGAGTGCGAAGGCGTGGTGTTAAA CGGTGAGTTTTGGGTTGTGAGTGGGTACGGCACCGAC **(SEQ ID NO:24)**

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4465	70904920	8,431373	A *	G	7,31	2,89	6,66	ATAGTGAGTTCTTTAATGAAACGTACCCATTCCATTATTTGGTTTAGTTTCATTT AATATTATTGAAAGTGGAACATCAAATTTAAGAAGATCCATATTTTAATTGGTTT TGGAAATTGTACATGAGGAATGATTAAGAAGACGTAGGCAACCAATTATAGAT GCACATTAATTTAGTTCATTTAGATCCTAATCCATGCA[A/G]CTGGCCCAATTAT TGGAGATCTCATGGTTTTGGGTATCTCAAATCTAAGCATAGCCACTGGCTCAG TCGGAGATCTCATGGTTCTTGCAATATCTTTGAGAATAAATTTCTGAATTTTCC CAGTTGAGGTCTTGGGAAGTTCATCTCTGAACACAACAACTTTAGGAACCATA TAATGAGGCAACTTCTCTCTACAATACT **(SEQ ID NO:25)**
common_4475	71862900	8,379069	A *	G	7,2	3,15	6,76	AGGTTACATCATCAATAAATGAATAGCACTACCAATAGAAGGAAAGAGTGATC ATGTGACATATTTCATATATGGCTCTTGAAACTAACCAAACACCACTCAAAAAA AATCTACATTTCTCTCCACATATCACACCTTCCCCTACTCATATTTTCTTCTATA TATGTCACTAAATCATAAACTCCCACATACAAGTACTCT[A/G]CTCAAGCCTTAT TACAAACACTCCTCTCTTCTACAAAAACCTCTCTGTTTCAAACATGGCTATAAG ATTACATGGAATGTTGAATGCTAAGCACCTCCTACGCCGCTCGAGTTTCTCTC AAAACCATGGAGTTTCAACACCCATCGACGTTCCCAAAGGCTATCTCGCGGT GTATGTTGGAGAAGAACACAAGAAGAGACA **(SEQ ID NO:26)**
common_4414	64950520	8,173088	A	G *	NA	6,8	5,66	CCTATTCAAGATATCATGGCCTTCAGTTCGGCAATTCATAATTCTTCGAAAGAC TTCTCTTAGTTGACGCTCCAATGGATCCAATCCATTTCGAAAACTTTCACAAAC CGCAGCTACCTCATGCACCCTTTGCTCTACCTCCTCCTTTTGCTCTTCTGTCA AGGGAAACTGAACAGAATCCACCAAGTCAGTCATGTGGCG[T/C]GCACATCTT TCAACCTGATAAATCTCCCTCAACAATCCATTGGTGTTTCGATTCTCTCTCTTT TTGGATTCTTCCAAAATCCGATTGAGAAGTGACATCAAAGGCGTTCCCCATGA GAAATTGCGGGGCACTGAGAAATGGACATTAAGACCGCGATCCTGACATGGG ATTGCAGCCACAAGAGTCCACAAAACAAACATG **(SEQ ID NO:27)**
common_4463	70588691	8,14361	A	G *	2,89	7,32	6,58	TGGTCAGTGTGAATTCCACTATTGTGTGAATTGTGTCACTTGCACATCCAGGC AAAGAGTAAAATATGAATATCATGATCATCCTCTTTTTTTTCTCGACAAAATTCA CACTGGTCTGGTGGAATGTGATGATTATGATTCTTATTGTAAGAATCCAATTTT GGCTGAATGTGACGAATTCAAGTATACCAATTCATCTGT[G/A]TTTGGTTGTGA TCAAGAATGTAATTTTAAAGTTCATTTGCTATGTGGTCCATTACCTAGCACTCT TAAATATGAATATCACATACATCCTCTTATTTTGTTTGATTTTGTTCTCAACAAT GATTATGGAGACTTTTACTGTGATATTTGTGAAATGAAAAGAGATCCACGAATA CGTGTATACTTTTGTGAAGATTGCAAC **(SEQ ID NO:28)**

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4504	74637355	8,033517	A *	G	7,48	2,36	6,66	TTAACTTAATGATCATATAGATAGTAATAAACTATTAATTAATTAATTTT GCTGGGGATGAGAATGGTGGCCAGGTAGCTTTTCCTTGATCTTTTCC ATAACAGTTTTCTTCTCGTGAAGAGGAGTAGGATCATGAGTAGTAGC AGTTGGTTCGTCCTTGTGTTTTCCAGTAAGTTTCTCTTTTATTTTTGTT CCTAAACC[C/T]TTCTTCTTCCTTCTCCCTCCAGCTCCATCATCCTCTG ACTGCATCCACACCCATTTTATAATTATTATATATATTAATCATTCATTT TTAATATATATAATTACATTATATTATTATTTTAATATTTTAAAATGTAAA TAAAAAATTTAGTGATTCAAATTATAATTTTTTATATTATATGTTGATAG CAATTTTGTTATT **(SEQ ID NO:29)**
rare_547	69138750	7,817705	A	C *	3,22	7,34	6,66	AAGACAAAAGAATAAGCTAATTAAGAAGCTAAAGAAAGAGATCAACTC TACTTTGAATTTGTTGTTTCAGAACAATATGACTACAAATCGATGGTG GGCTGGTAATGTGGCCATGAGAGGAGTGGATTCCATGTCTTCACCAC CACCACCACCTCCATTGCTTCAGCTCAGAAACACAGATGAAGATCCC AACAAAGATGA[C/A]GACCAAGACAGTGGCAACGGCGGCGACGACGA CGACCCCAATTCAACCGGCCACGAAAGCTTCGGACTAGGAGGAAGC AGCAGCAACAACCGCAGGCCACGTGGCAGACCCCCAGGTTCCAAGA ACAAGCCAAAGCCTCCGGTAGTGATAACAAAAGAGAGCCCCAATGCT CTAAGAAGCCACGTTTTGGAAATCAGTAGC **(SEQ ID NO:30)**
GBScompat_common_869	71021186	7,666785	A	G *	3	7,39	6,58	TATCTCCAAATACCCATCTGGGTGCATAACTCCCACATCGCCCGTGT AAAACCATCCATCCTCCTTTATAGCTTTAGCAGTAGCCAGTTCATCTT TAAGGTAGCCCAACATGACGCAATTACCTCTTAACACCACTTCCCCC ATTGTGAAACCATCTCTCTTCACACTCAACCCCGAATTCGGATCAACA ACATCAACTTC[A/G]GTCATTGCTGCTGAATTCACTCCTTGGCGCGCC TTTAGCCGGGCGCGCTCAGTAGCTGGAAGAAGATTCCACTTCTTCTT CCAAGCGCAAGACACAACGAGTCCGCACACTTCTGTCAACCCGTAG CCATGACTAACTTTAAAACCGAGCGACTCAGTTCGGTGTAGAACCGC CGCGGGAGGTGGAGCTCCTGCGGTAAGG **(SEQ ID NO:31)**
common_4462	70351069	7,629437	G *	A	7,32	2,89	6,63	TGTTTATATTATGAGGAAAAGAGGTTTGGTCCCAAGCTTGGTGTCATA TAATTCTATTGTACATGGACTTAGCAAGGAAGGAGGCTGTATGAGGG CTTATCAGTTGTTAGAAGAAGGCATTGAATTCGGATATCGACCATCTG AGCATACTTACAAGGTCTTAGTAGAAGGTCTTTGCCGAGAAAATGAC CTTCACAAGGC[G/A]AAATTTGTCTTTCATGTTATGCTCAACAAGGAAG GAGTTGAGAGAACTAGATTTTATAACATATATCTTAGAGCTCTTTGTTT TATGGATAATGCTGCTACTGAGCTTTTGAATACACTTGTTTCTATGCT CCAAACTCAGTGTCAACCTGATGTGATCACTCTCAATACCATTATTAA AGGGTTTTGTAAGATGGGGAGA **(SEQ ID NO:32)**

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4486	72986295	7,529719	C	G *	2,57	7,25	6,72	TCAATTTTGTTACTGTTATCTTTAATGCTTTCAATTGAATAAAATTC AAGGGGTTAATATTATAGCAATTCTGCAGCGTCTCTACATATTGA TTATCCTATGCTGTTTAAGCTTCAGAATGACTCTGCAAATAAAGT CTCACACTGTGGTGTCTTGGAGTTTATTGCTGAGGAAGGCATGA TATATATGCCTTATTGGGTA[C/G]GGATTGGTTTTTTCTGCGAAAT ATTCACCTGTACAGAGTTTTCTGGTCCATTTTTATCTAGCTTTTGA CCCATGGTGATTTGCACTCATTAGTGTTTTGTTGGTTCTGTTTAT CTATTAGATGATGCAAAATATGCTCTTAGGAGAGGGAGACTTTGT GCGTGTGAAGAATGTGACTCTTCTGAAGGGGACATATGTTAAA **(SEQ ID NO:33)**
common_4472	71630810	7,443011	A *	G	7,28	2,96	6,72	ATTCATGAAAGAAAATTTCTAATAATATTCATTTGTATCATTCTTG GGGACTACATTTAGTTGTTAATCCCCAAGGAAAAATTAATTACAA GGTATATGGCATTCTCTGCCTAACAAATTGTAACTGTGAGTTTTG TTAGCTGTTATCTTCTACTGGTGTCTAGACTCATTTGATTAGGAG CTTAAGTGAGAAATAAGATC[G/A]ATGAAGGCATCTTCTGTGCAT GGAATTGTAAGAGCTCCCATTGGATGATCGAACCCGAATTCTTCT TCAGCTTGATTTAGAAAGTCTTGGAATGAAAGCTGGTTCAAGTAT GCCACAGGGATCACAAATCTCTTCATTATGCTCTCATCGCGAACA TAAACCGCCAAAAAGCCTTTAGGGATATCTTTTGTGCTTGAGAAA **(SEQ ID NO:34)**
common_4517	75907527	7,402701	A *	C	7,43	2,33	6,7	CACAGCATCTCCATTAGAAGGATTAAACTTATGAGTACTAGCATG AGCATACCCACTTGCAAACCTAAAAAGACCACTTCCACCAATCAC AGGCATTTCTCTAACCTTATCAAACACTTGGTTTCTACCAAGAAT GGTGAGAGTGCTCCCATTGTATTTCCCTTGAGTTATATGAAAATT CATAGCCATAATTAGGGCAAT[C/A]TCTTCTTGTGAAGCTAATCCA TAAAACCCTTGAGCTTTTCCTAGCAACTTTGAGCTTACTTCTGGC CCTTCTGTCAATGGATTGTCGATCATGCTAACCGCCCCGAACCC ACTTTTCGATGCATTGGCCGGTGGTTGGATTATTGCCATCGCGC TAGGGTTTTTGCCGCTGTATATGTCGTGCCAATAGAACCGAAAG TGG **(SEQ ID NO:35)**

SNP	BP	LOD	Allele1	Allele2	Homo_1	Homo_2	Hetero	Context Sequence
common_4474	71780297	7,270959	A *	G	7,33	2,91	6,62	ACGTAATGCTTTGATGTATTTTCTGTCATCTAAATTCGGAATACCATCTAGAAT ATTTTGACAGTCTTCCACATTACATGAACTAGTGATATTGCTTGATTTTCCATT GATTATCTTCTCCCTTGTATCAATATACTTACCCATACTTTCACTTTGTACAAG TGATGCCTCAGTTTTAGCAGCTATACTTTTTGCAAGTTC[C/T]TCCATCACATG TTCAATGATAGAATTAGTCCCTTTGTTCTTTTTCTCCTTTGGTCTTGGTGCATC AAGTGGAACAGAAGTAGCTCGACGACGTACACCAGTAACTTCTTCTTCCTCT CCAAAATTTTCACCATCGCATTCAATACCATCATCAACATGTGCTCCAGTGTC AATATAATGCCGCTCTTGTTCATGTTCCTCA **(SEQ ID NO:36)**
common_4432	66679345	7,165743	A *	G	7,27	3,29	6,63	AGTAGTAGCAACGGTGAATTGACAACAAAAGTTTGAAGTCATTAACTACCCCA TGGCCATCAAAAACTCACCCTCACATACAACCTCACCTCCAACATATGCTTTC CCCTCCATTTTCGCTATTCCAAAGCGTTTCTGCAGCTTTGTGAGTGTCATTCT CATAACTAAAGTGTCACCAGCAATCACTGGTTTCCGGAATCT[A/G]ACTTTGTC GATTCCAGCAAAAAAGAAGTTCTCCCGAGAACCTCCCACTTCTGGTTGCAGC ATCACCAAACCACCAACCTGAGCCATTGCCTAGATCCAACAAAATTGAATATG TTGTTTTGTTAAAATTACATTACATGATGACTAAATAGATCAGAATATTTCCTA CAAGGACAGCAAAAAAATTCATTTTTCTAGGAAC **(SEQ ID NO:37)**
common_4452	69326994	7,113516	A *	G	7,33	3,18	6,62	GAGTGCCGTATATTTGTATTTAAACATTAGTCAACCAATATGATCAAATGTATA TATACGGTTACATATTACGCATATATATGAATCAAAGTTATATTACTTTCTCAAT ATGATCAAAGTTGTGATTTTGTTGGTGCTAGCCACACTCATTAATCAAGTATG GAGTAGTACTACAAGTAATAATTATTGCATAGAGAAGGA[G/A]AGACAAGCTC TTCTCAACTTGAAGAAAGGCTTTGTCGATGATGGCAATCGTCTATCCTCATGG ACAAGTAGTAGCCGTGATTGTTGTGCATGGAGAGGTATCAGGTGCGATAACT CAAAAACTCATCGTCATATTATCGCTCTTGATCTTAAATCTGATGACAACAATC ATAATTATTTGGGTGGTGAAATTGGTCCTTCT **(SEQ ID NO:8)**

EXAMPLE 4

Validation of purple color markers in Cannabis

The inventors identified SNP markers that are associated with purple color in whole cannabis plants. To validate the usefulness of the SNP markers identified they evaluated their effectiveness in predicting the presence of purple color in cannabis plants in a different F_2 population of cannabis plants. This F_2 population designated GID: 21 002 046 0000 was made from the selfing of a progeny of parents GID: 20 000 006 0000 known to not display the appearance of purple color in the whole plant and GID: 20 000 083 0000 known to display purple color in the whole plant.

Purple color was visually assessed of the whole plant from F_2 population GID: 21 002 046 0000 consisting of 113 individuals. At the time of harvest, plants were visually assessed for the presence of purple in the whole plant, the areas on leaf, stem, and flower. Plants were assessed for purple color in the whole plant with a score from 1 to 9, where 1 indicates a completely green plant and 9 a completely purple plant. A total of 30 (26.54% of total population) plants were scored less than 5 (more green), while 83 (73.45 % of total population) plants were scored greater than or equal to 5 (more purple). This indicates a dominant allele controlling purple color in the whole plant and the flower and that the trait is transmissible.

DNA was extracted from about 70 mg of leaf discs from all the plants evaluated using an adapted “sbeadex kit” with magnetic beads by LGC Genomics, automated on a KingFisher Flex with 96 Deep-Well Head by Thermo Fisher Scientific.

The extracted DNA served as a template for the subsequent library preparation for sequencing. The library pools were prepared according to the manufacturer’s instructions (AgriSeq™ HTS Library Kit—96 sample procedure from Thermo Fisher Scientific). Targeted sequencing of a custom SNP marker panel based on the Cannabis Sativa CS10 reference genome was carried out on the Ion Torrent system by Thermo Fisher Scientific. The primers for the SNPs identified are provided in Table 5. The library pool was loaded onto Ion 550 chips with Ion Chef and sequenced with Ion GeneStudio S5 Plus according to the manufacturer’s instructions (Ion 550™ Kit from Thermo Fisher Scientific).

From a population of 113 individuals, a genome-wide association analysis (GWAS) was performed to detect significant associations between genotypic information derived from targeted resequencing of the custom SNP marker panel described above and the appearance of purple color in the whole plant. Plants were assigned a score from 1 to 9, where 1 indicates a completely green plant and 9 a completely purple plant.

The genotypic matrix was filtered for SNPs having more than 30% missing values within the population and a minor allele frequency lower than 5 %. This resulted in 4015 SNP markers after filtering. The GWAS was performed using GAPIT version 3 (J. Wang & Zhang, 2021) with five statistical models: General Linear Model (GLM), Mixed Linear Model (MLM), FarmCPU and Blink (model=c("GLM", "MLM", "FarmCPU", "Blink"). A quantile-quantile plot (QQ plot) was used

to evaluate the statistical models. The Blink model performed the best by our evaluation and was used for the analysis. SNPs surpassing a LOD ($-\log_{10}$(p-value)) value of 5 were considered to have a significant association with trait variation.

Here the inventors identified additional SNPs significantly associate with purple color in the whole plant on chromosome NC_044377.1 and NC_044374.1 listed in Table 4 (Figure 2).

The inventors then looked specifically at the three SNPs on chromosome NC_044377.1 identified in Example 3 from population GID: 21 002 035 0000 with the highest LOD scores – "common_4519", "common_4525" and "common_4500" (Table 3). They found that in the F_2 population designated GID: 21 002 046 0000 these SNP markers were strongly linked to the gene and/or causative SNP underlying the appearance of purple color in cannabis, based on their LOD scores. These SNP markers can be used to predict of the presence or absence of purple color in the whole plant, including the flower. In the case of SNP "common_4519", "common_4525", and "common_4500" when homozygous for the indicative allele, in these cases at Allele 2, on average plants were found to have a score above 7, indicating a highly purple plant. For these three SNPs when heterozygous for the indicative allele, on average plants were found to have a score under 7, indicating that these were slightly less purple than the homozygous case. When, alternatively, the plants were homozygous for allele 1, on average the purple score was below 3.65, indicating a plant that is not purple. This shows that it is possible to produce a plant with purple color from a cross between a non-purple plant and a plant that is homozygous or heterozygous for the alleles associated with purple color without relying on the appearance of purple color to determine selection. Through the use of the provided markers it is possible to determine the allele state of the SNPs for the purple trait in order to identify the presence of the trait in the absence of the appearance of purple color to aid selection and identification of plants with with SNPs associated with the purple trait.

When considering the SNPs associated with purple color found in the GWAS from the two F_2 populations, a well-defined QTL on chromosome NC_044377.1 can be defined (Table 4, Figure 2). This QTL is well defined by the SNPs "GBScompat_common_864" and "GBScompat_common_879" at reference positions 68717484 to 77040783 on chromosome NC_044377.1. The SNP markers as well as the entire region that make up the QTL are linked to the gene and/or causative SNP underlying the appearance of purple color in cannabis as demonstrated by the linkage decay observed to a level under the LOD threshold of 5.

A second QTL associated with purple color can also be defined based on this experiment on NC_044374.1 based on the SNP markers "common_2448", "GBScompat_common_473", and "GBScompat_rare_86". This QTL is defined by the genomic region linked to these SNP markers and can be considered to be centered at position 6600328 on NC_044374.1 with reference to the CS10 genome of Cannabis Sativa.

Table 4: Validation of Purple Color in Whole Plant, F2 population 21 002 046 0000 showing the SNPs associated with the purple color trait. The presence of the purple color trait is predicted by the occurrence of the indicative allele (marked with *). The positions of the SNPs are provided with reference to the CS10 reference genome as described herein. Homo_1" denotes the average phenotypic value associated with homozygous allele 1 based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant, "Homo_2" denotes the average phenotypic value associated with homozygous allele based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant and "Hetero" denotes the average phenotypic value associated with heterozygous based on a score from 1-9, as described in the text, where 1 indicates a green plant and 9 indicates a purple plant. BP refers to the nucleotide position of the SNP.

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_4519	NC_044377.1	76201790	9,201071	A	G *	3,34	7,27	6,8	CGATCACTTCGTAGATGCATCCTCCCACAAGGTAGCACAATTG TAGAAAGTGCTAAATCATGCTTTATTCCATTTGTTCTTTTTGTCT TCTCTTTTTGCTTAATCGAACGATGTTGTGAACTTGTAGGGTTG TCAAATTCGAAGGGAGAGCGCACATGGAGAGAGCGTTTGTTG CAACAATGTGAGGGCCCTGTTTGATGA[A/G]CTCCCAACTCCAC ACCTAATTGTGGAGATCACACCATTCCCTGAAGGGCCTCTCAC TGAAAAAGATTACACCAAAGCTGAGAAATTGGAGAGGGTACTT AGAACTGGCCCGAACGTTTGATTCTTCTCTCGAGTTAAATCATC GCTGTCTCTCGTTAGAACTACAGCTTAATTGTATGTATGTTTTG AGCCTTGTACATAT **(SEQ ID NO:15)**
common_4526	NC_044377.1	76803154	8,582297	A	G *	3,61	7,38	6,76	TTTATATTACCTACATGAAATCAGACAACAACAGCATATCGGGG GCTTCACTCATGGAGATGGAGACCTACTACCATTTGGTTCATCT GGGTAGCGATTTGGACTTCGGCTAAGGCTTCTCTTTGGGCTTC GGCTTCTGCTTCGACTATGGCTTGGACTTCTAGCTCCACTTGG ACCCCTAGCTGGGCTTCGACTTGGGCT[T/C]CTTGATCCATTAT ATGGTGGAGAACGAGAGTACGACCTGTCTCGGCTGTACCTTCT CTCCTGTGGAGACACAGATCTGGCCCGAAGAGGGAGTACGAA AATTAATGTTTAAACATTGAAATATATTAAATGCATGTTTTCTCC TATTGCTAAAGATCCCACATTTTTAATGCTGACTAGAGAAGTTG AAAAGATATACTTG **(SEQ ID NO:38)**
common_4525	NC_044377.1	76757669	8,534953	A	C *	3,62	7,22	6,72	TTTATATTACCTACATGAAATCAGACAACAACAGCATATCGGGG GCTTCACTCATGGAGATGGAGACCTACTACCATTTGGTTCATCT GGGTAGCGATTTGGACTTCGGCTAAGGCTTCTCTTTGGGCTTC GGCTTCTGCTTCGACTATGGCTTGGACTTCTAGCTCCACTTGG ACCCCTAGCTGGGCTTCGACTTGGGCT[T/C]CTTGATCCATTAT ATGGTGGAGAACGAGAGTACGACCTGTCTCGGCTGTACCTTCT CTCCTGTGGAGACACAGATCTGGCCCGAAGAGGGAGTACGAA AATTAATGTTTAAACATTGAAATATATTAAATGCATGTTTTCTCC TATTGCTAAAGATCCCACATTTTTAATGCTGACTAGAGAAGTTG AAAAGATATACTTG **(SEQ ID NO:20)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
GBScompat_ common_879	NC_044377.1	77040783	8,407491	A *	T	7,33	3,82	6,8	CAAAATTATTTTAGTAAAGGTTCCTTCTTTAGATAAGAAAGAA CAAAATATGGCCATTGTATCATCAACCATTTCAAATAGGAAAC AATTAGATTCTCAGTACTCTAAACCTTTGATTCCTGATCCTGC TAGTACCAAACACATGGCTGCTCTTAAGCTTCAAAAGGTCTA CAAAAGCTTCCGAACTAGACGAAAGTTAGC[T/A]GATTGTGCT GTTCTTGTAGAGCAGAGTTGGTATGTAACTAATCTTTTATGTT CAATTCTAGTCTTTGTTGTGTTGTGTTGTTCTGTCATGTTCTG ATATACAATTTTTTGCATCAATTTTAGGTGGAAGCTCTTAGATT TTGCTGAACTCAAGAGGAGTTCGATATCTTTTTTCGATATTGA GAAGCATGAAACCGCGATT **(SEQ ID NO:39)**
common_4528	NC_044377.1	76959457	8,296782	A *	G	7,22	3,82	6,64	AAGTAAATAATATTTACATAAGTGAATTAGAATGAAGTAATAA GCAATAAAGTGCCACAAACTCCAACAAATCCAACTACAAATC CCACTCCAATTCCCATATAAAGCCATGACATGTTAAGCCACC CTTCATCATATTCTTCAGCATCACTTGGATCATGAATATCTTTA GGGCTATTTGGTGTCTCATCTCCAAGACAT[A/G]AATTGTTTA GTGGAAGTCCGCACAATCCATCATTATCAATGTATATCGAAG CATTAAAACTTTGCAATTGAGTACCGATAGGAATTCTTCCAGA CAATTTGTTACTTGACAAATTCAAAAAAGATAGAGAAGATATA CTTGCCAAGCTTGTTGGAATGACACTAGAAAGCTTGTTATGG GATAAATCTAGAGAATCCAACT **(SEQ ID NO:40)**
common_4518	NC_044377.1	75977377	7,797678	A *	G	7,3	3,17	6,87	TTTTATTTCTCTCACTTTTTTTGCATACTCTTTTCTTTCCATTCT TCTCGATCGTGCTGAAATACCTAATAAGACAGTGACACAGCA TGGCATGACACAATTAATGGGAGCGGTTGCCTTTGCACAACA ACTCCTCCTCTTCCACCTCCACTCTGCTGATCATATGGGACC AGAGGGACAATATCACTTGCTACTCCAGCT[C/T]GTGATTCTT GTCTCTCTGGTCACATCTCTAATGGGAATAGGGCTACCGAAG AGTTTCTTAGTGAGTTTTGTTAGGTCTCTTAGCATTTTGTTTCA AGGGGTTTGGCTTATGGTGATGGGGTTTATGTTATGGACACC ATCCTTGATTTCCAAAGGGTGTTTCATGCACTATGAGGAAGG TCATCATGTGGTGAGATGCTCA **(SEQ ID NO:41)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_4517	NC_044377.1	75907527	7,761131	A *	C	7,38	3,13	6,86	CACAGCATCTCCATTAGAAGGATTAAACTTATGAGTACTAGCATG AGCATACCCACTTGCAAACCTAAAAAGACCACTTCCACCAATCAC AGGCATTTCTCTAACCTTATCAAACACTTGGTTTCTACCAAGAAT GGTGAGAGTGCTCCCATTGTATTTCCCTTGAGTTATATGAAAATT CATAGCCATAATTAGGGCAAT[C/A]TCTTCTTGTGAAGCTAATCCA TAAAACCCTTGAGCTTTTCCTAGCAACTTTGAGCTTACTTCTGGC CCTTCTGTCAATGGATTGTCGATCATGCTAACCGCCCCGAACCC ACTTTTCGATGCATTGGCCGGTGGTTGGATTATTGCCATCGCGC TAGGGTTTTTGCCGCTGTATATGTCGTGCCAATAGAACCGAAAG TGG **(SEQ ID NO:35)**
GBScompat_ rare_164	NC_044377.1	72805722	7,710366	A *	C	7,15	3,77	6,83	AGCAGAGAAAATTGAAAAAATGACGGAAAGGAAAAAGAAGAGGG AAGCAACACCAGTATGCTTTGTGGCCACAGGGAAGTGAACAAGG TATAAATCCAGATAATCTAACTGAAGCTTCTTCAGACTATTCTTAC AGGCCTCAAGGACATGTCCATGGTCAGAATTCCAAAGCTGCAAT CAAGACAGGTAAAGAAAAAGAAT[A/C]ATAGTCTGGCTGGACATA CAGAATATACATATTATTCGATCAGTTCGAAAATAAACTAACCTTA GTGGTAACAAAGAGATCTTCTCTCTTAACAAGCCCTGTCTGAAAT GCCTCAGAAAGTGCCTCCCCAACTTCTGTTTCATTCCTGTAATCA GCTGCAACCAACACCAGCTAGTCAAACATTACATTGTAGCAACTC AA **(SEQ ID NO:22)**
common_4500	NC_044377.1	74383124	7,570859	A *	C	7,16	3,57	6,91	TCGACTTCAATGAGAAATCTCAACCCCTGTGGTAAGAATTTTGTG TACTTTTTTAAAAATTGAATTTTTAAAATTTCCCAGTTCGCTGAATA TTTGTATTAACTGTGTCTAATTTTTCATACTAGACATTGAAAGAAT GGTGTCTTTGAAGGGAATGGTTATCCGTTGCAGTTCAATAATTCC GGAAATCAGAGAAGCAAT[C/A]TTTAGATGTCTTGTGTGCGGTTA CTACTCTGAACCTGTTGCTGTAGAGAAAGGTAATTAGTTAGGATA CCATTCGCATGGGCTAGTTTTTTCTTATCTATATCACAATGTCATT CTAAATTATTTTCCCTTTTCAGGACGGATAACTGAACCAACAAGA TGCTTGAAGGAAGAATGCCAAGCAAGAAACTCCATGACACTT **(SEQ ID NO:13)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_2448	NC_044374.1	6600147	7,544585	A *	C	7,22	3,85	6,75	GAATGTGTTGGAATCCCATTCTTAATTTGATATTGAAATTTTT ATTGCTGTTTTCCAGTCTTGGCATGAAAGACTCTTGGGGGC CTTTGAAGGCTTTGGCTGTAGCTAGCATTATAAATGGCATT GGTGATATACTCCTGTGCAGAGTTTTTAGCTATGGCATTGC TGGTGCAGCATGGGCGACGATGGCATCACAGGTGC[T/G]C CAGATGAACATTTTGCTCCACTGTCTTTCCAGATTATATATT TCACTTTAGTTCTTATTAATTTCCGAGAAATATCCTAAATGA GTTTGTTTTCTTTCATACTGCCACTAACACAAATAGTATTAG GTTGTTGCAGGGTATATGATGGTTGAAAATCTGAACAAGAA AGGTTACAATGCTTATGCTCTCTCCATTCCCTC **(SEQ ID NO:42)**
GBScompat_ common_473	NC_044374.1	6600328	7,528498	A *	T	7,23	3,69	6,58	GACGATGGCATCACAGGTGCTCCAGATGAACATTTTGCTCC ACTGTCTTTCCAGATTATATATTTCACTTTAGTTCTTATTAAT TTCCGAGAAATATCCTAAATGAGTTTGTTTTCTTTCATACTG CCACTAACACAAATAGTATTAGGTTGTTGCAGGGTATATGA TGGTTGAAAATCTGAACAAGAAAGGTTACAATGC[T/A]TATG CTCTCTCCATTCCCTCACCGAAAGAACTTATCGCTATACTTG AGCTTGCTGCTCCGGTATTCATCACTATGACTTCTAAGGTA AATATTACTCAGTTTTCCTTGAGCTTGGCTATAATCTTTCCTT AGTTTTCCTTCAAAAACTAAGGTGTTTATATCCTTAGGTGGC ATTCTATAGTCTCCTCATATATTTTGCTA **(SEQ ID NO:43)**
common_4459	NC_044377.1	69980258	7,36557	A *	C	7,42	3,77	6,61	TAGGCAAAGCATGTCAAGACTGGGGCTTCTTTATGGTAATT ATTAATTAAGCTTAATTAATATTAATACCTCTAAAACAATATT GATGTCATATTTGAAATATTGCTATTCTTTTATGATTAATATA TAGGTGATCAATCATGGTGTGGCAGAGAGATTAATGAGTGA AGTTTTAGAAGGGTGTAGAGGTTTTTTTGATCT[T/G]AGTGA AGAAGAGAAGCTTGTGTTTAAGGGTACACATGTTATGGACA CAATTAGGTATGGTACAAGCTTCAATGCATCGGTAGAGAAA GCTTTGTATTGGAGAGATTATCTTAAGGTTCTTGTTCCTCAG CACCATCCTCATCATTTTCATTTCCCTAATAACCCATCTGGG TTCAGGTAATATTATTCACACATAAAATTT **(SEQ ID NO:44)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
common_4463	NC_044377.1	70588691	7,235412	A	G *	3,86	7,38	6,66	TGGTCAGTGTGAATTCCACTATTGTGTGAATTGTGTCACT TGCACATCCAGGCAAAGAGTAAAATATGAATATCATGATC ATCCTCTTTTTTTTCTCGACAAAATTCACACTGGTCTGGTG GAATGTGATGATTATGATTCTTATTGTAAGAATCCAATTTT GGCTGAATGTGACGAATTCAAGTATACCAATTCATCTGT[G /A]TTTGGTTGTGATCAAGAATGTAATTTTAAAGTTCATTTG CTATGTGGTCCATTACCTAGCACTCTTAAATATGAATATCA CATACATCCTCTTATTTTGTTTGATTTTGTTCTCAACAATGA TTATGGAGACTTTTACTGTGATATTTGTGAAATGAAAAGAG ATCCACGAATACGTGTATACTTTTGTGAAGATTGCAAC **(SEQ ID NO:28)**
common_4451	NC_044377.1	69163028	7,124024	A *	C	7,18	3,72	6,64	AAATGGCACCGCATCCGAAACAACAAATATCCCAATCAAA CGAAAGAACTCACTTATTGCCTTATGACACCGTCTTGCTT CATCCCCACTCTGATCATCGCTACCCGATGCTCCGAAATA ACGCTTTCCAGCCACCATTCTCACCACCATGTTAAGAGTT AAATCTTCTAACCAACGACTCAACTCAACTACAACTTTACA [A/C]GACTTGTAGAGCTCTCTAATCCCTACTTCAACCTCTG AAATCCTCACTTGCTTCAACATCTCTAGACGGCGGTTAGA GAGGAGTTCTAACGTGGCGATCTTCCTCATTTCGCGCCAA AAAGGGCTATAAGGTGCGAAACCAAAGACTGCGTAGTTG TAGCCCATGTGCTTGGCTGCCACGGTTGTAGGGCGCGAG GCCAGC **(SEQ ID NO:10)**
common_4483	NC_044377.1	72698292	7,106316	A	G *	3,81	7,25	6,74	CGACTTCTTTGCCGAAGATAAGAAAAACCCTGATAGCAGT TTCGATGAATATTTCTACGATGATGACGAAAAGCCTCGCG AGGAGTGTGGTGTTGTGGGTATTTATGGCGACTCAGAGG CCTCTCGGCTCTGTTACTTGGCTCTTCACGCTCTCCAACA TCGTGGTCAAGAAGGGGCTGGAATTGTTGCTGTGAAAAA CGA[C/T]GTTCTTCAATCCGTTACAGGCGTTGGACTTGTCT CTGAAGTCTTTAGCCATTCAAAGCTCGATCAATTGCCTGG AGATTTGGCTATTGGCCATGTACGGTACTCTACTGCTGGG TCTTCTATGCTTAAAAATGTTCAACCTTTTGTTGCAGGGTA TAGATTTGGTTCAGTTGGTGTTGCACACAATGGCAATTTG GTAAAT **(SEQ ID NO:45)**

SNP	Chromosome	BP	LOD	Allele_1	Allele_2	Homo_1	Homo_2	Hetero	Context Sequence
GBScompat_rare_86	NC_044374.1	6600352	7,082896	A	G *	3,74	7,25	6,5	GATGAACATTTTGCTCCACTGTCTTTCCAGATTATATATTTCACT TTAGTTCTTATTAATTTCCGAGAAATATCCTAAATGAGTTTGTTT TCTTTCATACTGCCACTAACACAAATAGTATTAGGTTGTTGCAG GGTATATGATGGTTGAAAATCTGAACAAGAAAGGTTACAATGCT TATGCTCTCTCCATTCCCTCACC[G/A]AAAGAACTTATCGCTATA CTTGAGCTTGCTGCTCCGGTATTCATCACTATGACTTCTAAGGT AAATATTACTCAGTTTTCCTTGAGCTTGGCTATAATCTTTCCTTA GTTTTCCTTCAAAAACTAAGGTGTTTATATCCTTAGGTGGCATT CTATAGTCTCCTCATATATTTTGCTACATCCATGGGCACAATCA GCATGG **(SEQ ID NO:46)**
GBScompat_common_864	NC_044377.1	68717484	7,054155	A	G *	3,77	7,35	6,64	TAGTGGGTTAGGAACTGGGAGCAAAGGCCTAAGCGGCGGACA GAAGAGGAGAGTGAGCATCTGCATAGAGATTCTCACTCGTCCA AAGCTTCTTTTCCTTGACGAGCCCACTAGTGGGCTCGACAGTG CTGCTTCATACTATGTGATGAGCAGCATTGCGTCGTTGGATATT CAGAGTCGGAGGGGTGGTGGGGCCGGTGG[C/T]CGGAGGACT GTGGTGGCTTCCATCCACCAGCCCAGTTCCGAAGTGTTTCAGC TTTTTAATACTCTTTGCCTTCTTTCTGCTGGTAAAATTGTGTATT TTGGTCCTGCTAGTGCAGCTAATGAGGTATTTTCAGGTTTTTTT TAAACGTATTTAAATTATAAATTACAATACATAAAGGAATAATAA TATTTGTACTAATTA **(SEQ ID NO:47)**

Table 5: Targeted sequencing primers for the SNPs identified in Tables 1 to 4, as described in Examples 1 to 4.

SNP	Forward Primer	SEQ ID NO	Reverse Primer	SEQ ID NO
common_1535	GCAAATGAGTTGGTGAGCCA	SEQ ID NO:48	GCCAATGCAGCATCCAACTC	SEQ ID NO:49
common_2032	GGGGTGATGTGAATGAGTGC	SEQ ID NO:50	CCTGGGATGCTGTGTGTTCT	SEQ ID NO:51
common_2262	AAATTAAGAGCAGTTTCACGTGT	SEQ ID NO:52	TGCTCGAATCTGCTACATTGC	SEQ ID NO:53
common_2448	AAGACTCTTGGGGGCCTTTG	SEQ ID NO:54	CCATCATATACCCTGCAACAACC	SEQ ID NO:55
common_4054	TTACCCATCGTGGCTGACTG	SEQ ID NO:56	CGCAACTAACGGCAACCAAA	SEQ ID NO:57
common_4414	TTCACAAACCGCAGCTACCT	SEQ ID NO:58	TGGACTCTTGTGGCTGCAAT	SEQ ID NO:59
common_4432	AGCAACGGTGAATTGACAACA	SEQ ID NO:60	AATGGCTCAGGTTGGTGGTT	SEQ ID NO:61
common_4446	TTGGAAGAATGGCGGGTCAG	SEQ ID NO:62	TATCCCCCTGCTCACTGTGA	SEQ ID NO:63
common_4451	ACACCGTCTTGCTTCATCCC	SEQ ID NO:64	TGGGCTACAACTACGCAGTC	SEQ ID NO:65
common_4452	TGTTGGTGCTAGCCACACTC	SEQ ID NO:66	AGGACCAATTTCACCACCCA	SEQ ID NO:67
common_4459	AAGCATGTCAAGACTGGGGC	SEQ ID NO:68	ACCGATGCATTGAAGCTTGT	SEQ ID NO:69
common_4462	GAAGGAGGCTGTATGAGGGC	SEQ ID NO:70	GGTTGACACTGAGTTTGGAGC	SEQ ID NO:71
common_4465	AAGGGTAAGGGAGAGTGGCA	SEQ ID NO:72	AGCTCGACTAATGGGGATGA	SEQ ID NO:73
common_4472	GCATTCTCTGCCTAACAAATTGT	SEQ ID NO:74	CCTAAAGGCTTTTTGGCGGT	SEQ ID NO:75
common_4474	GTGATGCCTCAGTTTTAGCAGC	SEQ ID NO:76	GGAACATGAACAAGAGCGGC	SEQ ID NO:77
common_4475	TCCACATATCACACCTTCCCC	SEQ ID NO:78	TCTCCAACATACACCGCGAG	SEQ ID NO:79
common_4483	TTATGGCGACTCAGAGGCCT	SEQ ID NO:80	ACCAAATTGCCATTGTGTGCA	SEQ ID NO:81
common_4485	TATTTTCCGGCGTTCCTCCG	SEQ ID NO:82	GGAGAGACTCAAGGCGGTTC	SEQ ID NO:83
common_4486	CAATTCTGCAGCGTCTCTACA	SEQ ID NO:84	CACGCACAAAGTCTCCCTCT	SEQ ID NO:85
common_4487	TGTTTCGTGGGTCGTTTTCC	SEQ ID NO:86	ACCTTGTCATCGCAGGAACT	SEQ ID NO:87
common_4499	CAGATAGGACACTCTGAACAAGC	SEQ ID NO:88	TCAGGCTCGAATGCAAGAGA	SEQ ID NO:89
common_4500	GAGAAATCTCAACCCCTGTGGT	SEQ ID NO:90	AACTAGCCCATGCGAATGGT	SEQ ID NO:91
common_4502	TCAGTGGAGTCTCTGGTGACA	SEQ ID NO:92	GGCTAGCCCGCTTTATCACA	SEQ ID NO:93
common_4504	TGCTGGGGATGAGAATGGTG	SEQ ID NO:94	ATGGGTGTGGATGCAGTCAG	SEQ ID NO:95
common_4513	AGAGAAGGCCATCTCCGAGT	SEQ ID NO:96	GAATCTTACGAGGCAACGAGC	SEQ ID NO:97
common_4514	ACCTCGGAGACCATGTCATTG	SEQ ID NO:98	GGTTTGGTTGGGGAAACAGG	SEQ ID NO:99
common_4517	GCATGAGCATACCCACTTGC	SEQ ID NO:100	CGGCCAATGCATCGAAAAGT	SEQ ID NO:101
common_4518	TTAATGGGAGCGGTTGCCTT	SEQ ID NO:102	GCATCTCACCACATGATGACC	SEQ ID NO:103
common_4519	GATGCATCCTCCCACAAGGT	SEQ ID NO:104	CGGGCCAGTTCTAAGTACCC	SEQ ID NO:105
common_4522	GGACGGGTTGCCTATGTTCT	SEQ ID NO:106	ACTCATCTGAGCCAACTCGC	SEQ ID NO:107
common_4525	CAAGTGCTTGCTGAGGATGG	SEQ ID NO:108	GCGCTACAGGCTCTCAGAAT	SEQ ID NO:109

common_4526	CAACAGCATATCGGGGGCTT	SEQ ID NO:110	CTCTTCGGGCCAGATCTGTG	SEQ ID NO:111
common_4528	ACAAATCCCACTCCAATTCCCA	SEQ ID NO:112	GTCATTCCAACAAGCTTGGCA	SEQ ID NO:113
common_4599	ACGAAAAAGAGGCGCTTCTTG	SEQ ID NO:114	AGGTCCAGTTAAGCTGTTGTAAGT	SEQ ID NO:115
common_5220	ATTGATGGTGGTGATGGGCA	SEQ ID NO:116	TCCGAGGACGACTTTCTAGGT	SEQ ID NO:117
common_816	CCTGATCATGGGCAGCATCA	SEQ ID NO:118	TTAGAAACAGCCTGGTGGGG	SEQ ID NO:119
GBScompat_ common_473	GCATCACAGGTGCTCCAGAT	SEQ ID NO:120	AGTGATGAATACCGGAGCAGC	SEQ ID NO:121
GBScompat_ common_864	GTGGGTTAGGAACTGGGAGC	SEQ ID NO:122	AAACACTTCGGAACTGGGCT	SEQ ID NO:123
GBScompat_ common_869	ACATCGCCCGTGTAAAACCA	SEQ ID NO:124	TCATGGCTACGGGTTGACAG	SEQ ID NO:125
GBScompat_ common_879	ACCAAACACATGGCTGCTCT	SEQ ID NO:126	ATCGCGGTTTCATGCTTCTC	SEQ ID NO:127
GBScompat_ rare_164	GGGAAGCAACACCAGTATGC	SEQ ID NO:128	AACAGAAGTTGGGGAGGCAC	SEQ ID NO:129
GBScompat_ rare_165	CAGCCCAAACCTTTTGAGCA	SEQ ID NO:130	CAGGCGTACTGTTGTGAGCA	SEQ ID NO:131
GBScompat_ rare_2	ACAAACTGCTGCCTCTGTATCT	SEQ ID NO:132	GTGCCAGCCATTCTCAAAGC	SEQ ID NO:133
GBScompat_ rare_86	TGCTCCACTGTCTTTCCAGA	SEQ ID NO:134	GCCAAGCTCAAGGAAAACTGA	SEQ ID NO:135
rare_547	TCGATGGTGGGCTGGTAATG	SEQ ID NO:136	TGGCTTCTTAGAGCATTGGGG	SEQ ID NO:137
common_4448	TGATCAGCGAAGAAAGGCCA	SEQ ID NO:138	AGCATCACGGCTATGACACC	SEQ ID NO:139
common_4463	TCACTTGCACATCCAGGCAA	SEQ ID NO:140	AGTCTCCATAATCATTGTTGAGAACA	SEQ ID NO:141

EXAMPLE 5

Gene Identification

There are presently no known genes identified in Cannabis that have been shown to regulate color in flowers or throughout the whole plant. Genes that regulate flower color through the biosynthesis of anthocyanins or through their transcriptional regulation have been described and characterized in several plant species. The inventors considered genes that regulate anthocyanin levels as being the best candidate genes for controlling the appearance of purple in cannabis color observed. They next sought to identify putative genes that could encode proteins that may be responsible for the accumulation of anthocyanins in the total plant and in the flower. Using the findings of the association studies they identified candidate genes at the QTLs identified.

At the QTL found on chromosome NC_044373.1 based on the SNP “common_2262” at position 80922439, the inventors looked at a 2mB region centering on this SNP for putative candidate genes. Genes and annotation of the reference genome CS10 (GCF_900626175.1) were retrieved from NCBI. Scans for known amino acid domains were performed using hmmer (v3.1, http://hmmer.janelia.org/, with the option -E 1e-5) with the Pfam database (v33, Finn et al). Gene description, related KEGG pathways and GO terms were identified using Pannzer (v2, Toronen et al.) using default settings and manually inspected. The inventors identified two genes with gene ID LOC115712034 and LOC115712567 listed in Table 6. Both are annotated as acyl-transferase family

proteins. A BLAST search of the amino acid sequences encoded by these genes of all Arabidopsis thaliana proteins returned an HXXXD-type acyl-transferase family protein as the closest homolog. Acyl-transferases, like the two identified, may be involved in transferring acyl-groups to the sugar moieties of anthocyanins affecting the purple color of plant tissue through the stability of the anthocyanin, causing them to either accumulate or dissipate.

Based on the results of the association study for purple color from the F2 population 21 002 046 0000 the inventors identified a QTL on NC_044374.1 marked by the three SNPs – "common_2448", "GBScompat_common_473", and "GBScompat_rare_86". They looked for putative candidates in the region of this QTL by manual inspection of an annotated gene list for chromosome NC_044374.1 from the Cannabis sativa CS10 genome. The inventors identified a candidate gene within 0.1Mb that is annotated to encode an acyl transferase family protein, with gene ID LOC115716241 listed in Table 6. A BLAST search of the amino acid sequences encoded by these genes of all Arabidopsis thaliana proteins returned an HXXXD-type acyl-transferase family protein as the closest homolog. Acyltransferases like the two identified may be involved in transferring acyl-groups to the sugar moieties of anthocyanins affecting the purple color of plant tissue through the stability of the anthocyanin, causing them to either accumulate or dissipate.

From the QTL found on NC_044377.1 between position 64950520 – 77040783 the inventors searched for genes that may encode proteins involved in the biosynthesis or transcriptional regulation of anthocyanins from an annotated gene list for this region of NC_044377.1 from the Cannabis sativa CS10 genome. Upon inspection of this genomic region and BLAST analysis of putative candidates they identified five candidate genes LOC115695758, LOC115725215, LOC115695887, LOC115695872, LOC115695872 listed in Table 6. The gene IDs LOC115695758, LOC115725215, LOC115695887 encode putative MYB Transcription factors. MYB Transcription factors in other plant species act as regulators of secondary metabolism, including positively and negatively regulating anthocyanin biosynthesis.

The inventors also identify two genes LOC115725215 and LOC115695887 that are annotated as encoding putative anthocyanidin 3-O-glucosyltransferase. Glucosyltransferase proteins transfer the sugar moiety to anthocyanidin. Anthocyadins are stabilized by the addition of a sugar moiety. This suggests a mechanism for the regulation of purple color in cannabis whereby the loss or gain of function of this protein would affect the accumulation of anthocyanins in plant tissue.

Table 6: Gene list of candidate genes identified. The gene ID is provided with reference to the publicly available CS10 genome.

Chromosome	**Start Position**	**End Position**	**Gene ID**	**Protein ID**	**Description**
NC_044373.1	79836159	79837767	LOC115712034	XP_030496100.1	HXXXD-type acyl-transferase family protein
NC_044373.1	79968804	79970536	LOC115712567	XP_030496724.1	HXXXD-type acyl-transferase family protein
NC_044374.1	6676610	6680038	LOC115716241	XP_030500856.1	HXXXD-type acyl-transferase family protein
NC_044377.1	75409856	75419127	LOC115695758	XP_030478701.1	MYB domain TF, 2 SANT Domains
NC_044377.1	75894244	75898321	LOC115725215	XP_030510513.1	MYB domain TF
NC_044377.1	76275921	76280609	LOC115695887	XP_030478846.1	MYB domain TF
NC_044377.1	76403822	76405684	LOC115695872	XP_030478824.1	anthocyanidin 3-O-glucosyltransferase 2
NC_044377.1	76416328	76418041	LOC115695871	XP_030478823.1	anthocyanidin 3-O-glucosyltransferase 2

CLAIMS:

1. A method for identifying a Cannabis sativa plant comprising in its genome a purple color QTL, the method comprising the steps of:

(i) genotyping at least one plant with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4; and

(ii) identifying one or more plants containing the purple color QTL.

2. The method of claim 1, wherein the polymorphism is selected from the group consisting of "common_4519", "common_4525", and "common_4500", as defined in Table 4.

3. The method of claim 1 or 2, wherein the genotyping is performed by PCR-based detection using molecular markers, sequencing of PCR products containing the one or more polymorphisms, targeted resequencing, whole genome sequencing, or restriction-based methods, for detecting the one or more polymorphisms.

4. The method of claim 3, wherein the molecular markers are for detecting polymorphisms at regular intervals within the purple color QTL such that recombination can be excluded.

5. The method of claim 3, wherein the molecular markers are for detecting polymorphisms at regular intervals within the purple color QTL such that recombination can be quantified to estimate linkage disequilibrium between a particular polymorphism and a purple color phenotype.

6. The method of any one of claims 3 to 5, wherein the molecular markers are selected from the primer pairs as defined in Table 5.

7. A method of producing a Cannabis sativa plant having a purple color QTL in its genome, the method comprising the steps of:

(i) providing a donor parent plant having in its genome a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4;

(ii) crossing the donor parent plant having the purple color QTL with at least one recipient parent plant that does not have the purple color QTL to obtain a progeny population of cannabis plants;

(iii) screening the progeny population of cannabis plants for the presence of the purple color QTL; and

(iv) selecting one or more progeny plants having the purple color QTL, wherein the plant displays the purple color trait.

8. The method of claim 7, further comprising:

(v) crossing the one or more progeny plants with the donor recipient plant; or

(vi) selfing the one or more progeny plants.

9. The method of claim 7 or 8, wherein the screening comprises genotyping at least one plant from the progeny population with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4.

10. The method of any one of claims 7 to 9, wherein the method comprises a step of genotyping the donor parent plant with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4.

11. The method of any one of claims 7 to 10, wherein the polymorphism is selected from the group consisting of “common_4519”, “common_4525”, and “common_4500”, as defined in Table 4.

12. A method of producing a Cannabis sativa plant that does not include a purple color QTL in its genome, the method comprising the steps of:

(i) providing a donor parent plant having in its genome a QTL associated with an absence of purple color characterized by one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4;

(ii) crossing the donor parent plant having the QTL associated with the absence of purple color with at least one recipient parent plant that has a purple color QTL to obtain a progeny population of cannabis plants;

(iii) screening the progeny population of cannabis plants for the presence of the QTL associated with the absence of purple color; and

(iv) selecting one or more progeny plants having the QTL associated with the absence of purple color, wherein the plant does not display the purple color trait.

13. The method of claim 12, further comprising:

(v) crossing the one or more progeny plants with the donor recipient plant; or

(vi) selfing the one or more progeny plants.

14. The method of claim 12 or 13, wherein the screening comprises genotyping at least one plant from the progeny population with respect to the QTL associated with the absence of purple color by detecting one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4.

15. The method of any one of claims 12 to 14, wherein the method comprises a step of genotyping the donor parent plant with respect to the purple color QTL by detecting one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4.

16. The method of any one of claims 12 to 15, wherein the polymorphism is selected from the group consisting of "common_4519", "common_4525", and "common_4500", as defined in Table 4.

17. The method of any one of claims 9, 10, 14 and 15, wherein the genotyping is performed by PCR-based detection using molecular markers, sequencing of PCR products containing the one or more polymorphisms, targeted resequencing, whole genome sequencing, or restriction-based methods, for detecting the one or more polymorphisms.

18. The method of claim 17, wherein the molecular markers are for detecting polymorphisms at regular intervals within the QTL such that recombination can be excluded or such that recombination can be quantified to estimate linkage disequilibrium between a particular polymorphism and a purple color phenotype or absence of purple color phenotype.

19. A method of producing a Cannabis sativa plant comprising a purple color trait, the method comprising introducing a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4 into a Cannabis sativa plant, wherein said purple flower QTL is associated with the purple color trait.

20. The method of claim 19, wherein introducing the purple color QTL comprises crossing a donor parent plant in which the purple color QTL is present, with a recipient parent plant in which the purple color QTL is not present.

21. The method of claim 19, wherein introducing the purple color QTL comprises genetically modifying the Cannabis sativa plant.

22. A method of producing a Cannabis sativa plant that does not display a purple color trait, the method comprising introducing a QTL characterized by one or more polymorphisms

associated with the absence of purple color as defined in any one of Tables 1 to 4 into a Cannabis sativa plant, wherein said QTL is associated with the absence of purple color in the plant.

23. The method of claim 22, wherein introducing the QTL comprises crossing a donor parent plant in which the QTL associated with the absence of purple color is present, with a recipient parent plant in which the QTL is not present.

24. The method of claim 22, wherein introducing the QTL associated with the absence of purple color comprises genetically modifying the Cannabis sativa plant.

25. A Cannabis sativa plant identified according to the method of any one of claims 1 to 6, or produced according to the method of any one of claims 7 to 24, provided that the plant is not exclusively obtained by means of an essentially biological process.

26. A Cannabis sativa plant comprising a purple color QTL characterized by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4, provided that the plant is not exclusively obtained by means of an essentially biological process.

27. A Cannabis sativa plant comprising a QTL associated with the absence of purple color characterized by one or more polymorphisms associated with the absence of purple color as defined in any one of Tables 1 to 4, provided that the plant is not exclusively obtained by means of an essentially biological process.

28. A quantitative trait locus that controls a purple color trait in Cannabis sativa, wherein the quantitative trait locus is defined by a single nucleotide polymorphism at position 80922439 of NC_044373.1 or a genetic marker linked to the QTL.

29. A quantitative trait locus that controls a purple color trait in Cannabis sativa, wherein the quantitative trait locus is defined by a single nucleotide polymorphism at position 6600328 of NC_044374 or a genetic marker linked to the QTL.

30. A quantitative trait locus that controls a purple color trait in Cannabis sativa, wherein the quantitative trait locus has a sequence that corresponds to nucleotides 68717484 to 77040783 of NC_044377.1 and is defined by one or more polymorphisms associated with purple color as defined in any one of Tables 1 to 4 or a genetic marker linked to the QTL.

31. An isolated gene that controls a purple color trait in a Cannabis sativa plant, wherein the gene is selected from the group consisting of the genes as defined in Table 6 with reference to the CS10 genome.

32. The isolated gene of claim 31, wherein the gene has the gene identity number LOC115695758 and encodes a putative MYB Transcription factor.

33. The isolated gene of claim 31, wherein the gene has the gene identity number LOC115695872 or LOC115695871 and encodes an anthocyanidin 3-O-glucosyltransferase 2.

Intellectual Property Office

Application No:	GB2204468.9	**Examiner:**	Dr Jeremy Kaye
Claims searched:	1-30	**Date of search:**	25 May 2022

Patents Act 1977: Search Report under Section 17

Documents considered to be relevant:

Category	Relevant to claims	Identity of document and passage or figure of particular relevance
A	-	WO2020/093101 A1 (AGRIC. VICTORIA SERVICES PTY LTD) see para.[0190]
A	-	Genome Biol., Vol.12, 2011, van Bakel, H. et al., "The draft genome and transcriptome...", p.R102 Open Access: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-10-r102
A	-	J. Cannabis Res., Vol.1, 2019, Schwabe, A. L. & McGlaughlin, M. E., "Genetic tools weed out misconceptions...", Article No.: 3 Open Access: https://jcannabisresearch.biomedcentral.com/articles/10.1186/s42238-019-0001-1

Categories:

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.

Field of Search:

Search of GB, EP, WO & US patent documents classified in the following areas of the UKC^{X} :

Worldwide search of patent documents classified in the following areas of the IPC

The following online and other databases have been used in the preparation of this search report

WPI, EPODOC, Patent Fulltext, BIOSIS, MEDLINE

International Classification:

Subclass	Subgroup	Valid From
None		