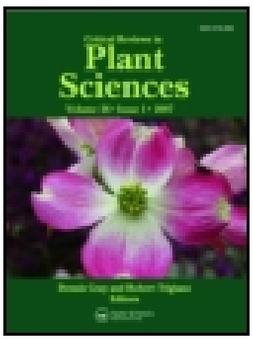

Critical Reviews in Plant Sciences

ISSN: 0735-2689 (Print) 1549-7836 (Online) Journal homepage: http://www.tandfonline.com/loi/bpts20

Genetic and Genomic Tools for Cannabis sativa

Daniela Vergara, Halie Baker, Kayla Clancy, Kyle G. Keepers, J. Paul Mendieta, Christopher S. Pauli, Silas B. Tittes, Kristin H. White & Nolan C. Kane

To cite this article: Daniela Vergara, Halie Baker, Kayla Clancy, Kyle G. Keepers, J. Paul Mendieta, Christopher S. Pauli, Silas B. Tittes, Kristin H. White & Nolan C. Kane (2017): Genetic and Genomic Tools for Cannabis sativa, Critical Reviews in Plant Sciences, DOI: 10.1080/07352689.2016.1267496

To link to this article: http://dx.doi.org/10.1080/07352689.2016.1267496

CRITICAL REVIEWS IN PLANT SCIENCES
http://dx.doi.org/10.1080/07352689.2016.1267496

Taylor & Francis
Taylor & Francis Group

Genetic and Genomic Tools for *Cannabis sativa*

Daniela Vergara, Halie Baker, Kayla Clancy, Kyle G. Keepers, J. Paul Mendieta, Christopher S. Pauli, Silas B. Tittes, Kristin H. White, and Nolan C. Kane

Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, USA

ABSTRACT

The *Cannabis* industry is currently one of the fastest growing industries in the United States. Given the changing legal status of the plant, and the rapidly advancing research, updated information on the advancement of *Cannabis* genomics is needed. This versatile plant is used as medicine and for food, fiber, and bioremediation. Insights from modern, high-throughput genomic technology are revolutionizing our understanding of the plant and are providing new tools to further improve our knowledge and utilization of this unique species. This review quantifies and evaluates the currently available genomic resources for *Cannabis* research, including six whole-genome assemblies, two transcriptomes, and 393 other substantial genomic resources, as well as other smaller publicly available genetic and genomic resources. The open-source approaches followed by many leading scientists in the field promote collaboration and facilitate these rapid advances.

KEYWORDS

Genetic map; genome assemblies; genomic resources; hemp; marijuana; sex chromosomes; transcriptomes

I. Introduction

Cannabis, a genus within the Rosales clade of eudicot angiosperms, is the basis for a multibillion-dollar industry, one of the fastest growing industries in the US. Yet, in many respects the plant remains poorly understood. The plant is cultivated for its fibrous stalks, and its seeds which are high in healthy oils and protein (Gertsch *et al.*, 2008; Zwenger and Basu, 2008), responsible for psychoactive effects (Russo and McPartland, 2003; ElSohly and Slade, 2005; Radwan *et al.*, 2008) and medicinal properties (Russo, 2011; Swift *et al.*, 2013; Volkow *et al.*, 2014). *Cannabis* has a tremendous amount of phenotypic diversity seen in traits such as leaf shape, plant shape, seed size, fiber quality, branching, height, phytochemistry, phenology, and ecology. Additionally, there is substantial variation in reproductive strategy, with some populations maintaining the ancestral, monoecious type, while others maintaining a more derived dioecious or mixed strategy (Razumova *et al.*, 2016). Some trait variations are related to the multiple domesticated uses, whereas additional variation in the wild is presumably related to ecological differences across its broad Eurasian range, reaching from Eastern Europe to Japan and from the equator to Siberia.

The number of species that compose the genus is debated; some claim that it is composed of three species (*Cannabis sativa, Cannabis indica*, and *Cannabis ruderalis*) (Hillig and Mahlberg, 2004; Hillig, 2005), whereas others claim that it is a monotypic genus composed of one species (*C. sativa* L.) with high interpopulation variation (Small and Cronquist, 1976; de Meijer *et al.*, 2003). Currently, according to most scientific conventions, the plant is classified as one species, *C. sativa*, originally described by Linnaeus (1753). Most of the groupings or divisions in this genus have been made depending on the physical traits or uses of the plant. Hemp plants are usually used for the extraction of fiber or oil, and the plants are described as tall, skinny, and lacking the production of terpenoids or cannabinoids (Small and Cronquist, 1976). *Indica* plants are usually short, with broad leaflets. Sedative effects are associated with its consumption. *Sativa* plants are associated with narrow leaflets and hemp-like traits, but unlike hemp produce high levels of cannabinoids and terpenoids. Psychedelic effects are associated with the consumption of *sativa*. Crosses between *sativa* and *indica* are common, vigorous, and fertile, and are generally referred to as *hybrid* within the *Cannabis* industry. These hybrids have variable phenotypic traits usually intermediate to the parents, though in both hemp and marijuana types, heterosis leads to improved

CONTACT Daniela Vergara daniela.vergara@colorado.edu; Nolan C. Kane nolan.kane@colorado.edu Department of Ecology and Evolutionary Biology, University of Colorado Boulder, 1900 Pleasant Street, Boulder, CO 80309, USA.

Color versions of one or more of the figures in this article can be found online at www.tandfonline.com/bpts.

The authors Halie Baker, Kayla Clancy, Kyle G. Keepers, J. Paul Mendieta, Christopher S. Pauli, Silas B. Tittes, and Kristin H. White contributed equally to this work.

growth and other traits that exceed either parental type (Clarke and Merlin, 2013).

Due to the multiple classifications of this plant, genomic studies provide insights on how to better group the individuals. Additionally, genetic and genomic resources available for *Cannabis* research have been increasing, and may become more widely used because of the ongoing legalization and decriminalization of its medicinal and recreational use in increasingly large number of states in the United States, as well as in other countries. In this paper, we review the available genomes, their sequencing methods, and the relationship among *Cannabis* cultivars; the available resources for genetic regions of interest including those involved in the production of cannabinoids and terpenoids; whole genome and cytoplasmic genome assemblies as well as transcriptomes; and the genomic underpinnings involving important agronomic traits, including sex determination and cannabinoid production. We conclude with future directions and suggestions on how to proceed with *Cannabis* genomic investigation, which will benefit scientific research, patients, breeders, growers, recreational consumers, and the general public.

Because of the multiple naming conventions, we will address the plant using its scientific genus *Cannabis*, but we will also use the colloquial terminologies for major lineages within the species, using *indica, sativa, hemp*, and *hybrids* in *italics* throughout the review.

II. Available genomic sequences

A. Genome assemblies

There are six whole-genome assemblies publicly available, consisting of two drafts, each of three different genotypes. Two Purple Kush (PK) assemblies (van Bakel *et al.*, 2011) are available in the National Center for Biotechnology Information (NCBI) (Accessions: AGQN00000000; GCA_000230575.1). They were assembled from 2 × 100 paired-end Illumina libraries; 2 and 5 kb Illumina mate-pair libraries; and 8, 13, 20, 30, and 40 kb Roche 454 mate-pair libraries. Sequences were assembled *de novo* using SOAPdenovo (Li *et al.*, 2010). The PK assembly with accession AGQN00000000 is superior to the GCA_000230575 assembly by the metrics we examined (Table 1), as it does not include the small, lower-quality, unplaced contigs.

Four additional genomic assemblies exist, consisting of sequences from two strains, with raw and filtered assemblies for each variety. The filtered version of Chemdog (Accession: GCA_001509995) was sequenced on the Illumina GA IIx and assembled in CLCBio v.5. The filtered LA Confidential (Accession: GCA_001510005) genome was sequenced on the Roche 454 and assembled using Newbler v.May 2011. The unfiltered versions are available at http://www.medicinalgenomics.com/janeome-additional-data/.

The PK scaffold assembly is the best assembly by most (eight of twelve) metrics (Table 1): the full sequence length requires fewer contigs and the average sequence length is higher. Additionally, the metrics for L50 and L90, which are the smallest number of contigs required to represent 50% and 90% of the assembly, respectively, are smaller for the PK assemblies. The N50 and N90 metrics are larger in the PK assemblies, again indicating that the PK assembly is made up of longer sequences. The unfiltered PK assembly is the best or tied for the best for the remaining four metrics. Therefore for research including the currently available genomic assemblies, we recommend the use of the PK assemblies.

In order to understand the completeness of the single-copy portion of the assemblies, we performed a BLAST comparison using NCBI's BLASTX algorithm (Altschul *et al.*, 1990; Gish and States, 1993) with the default parameters. This BLAST comparison allows us to identify in each assembly the DNA sequence encoding for a set of 956 universal single-copy plant orthologs, which are thought to be essential genes expected to be present in all high-quality, complete plant genomes (Simão *et al.*, 2015). We report the percentage of the number of these Benchmarking Universal Single-Copy Orthologs (BUSCO) genes that were uniquely present in the assemblies, a measure of an assembly's completeness. The results show that the number of genes in the single-copy portion of the assemblies is comparable between all the available assemblies. All of them are nearly complete in terms of having at least partial copies of all the BUSCO genes, but in contrast to the transcriptomes, where nearly all of the genes are complete, many of the genes in the genome assemblies are fragmented or incomplete (Table 1). A more completely assembled genome, ideally with 11 long sequences representing the 9 autosomes, X and Y, would be a major improvement.

In addition to the above metrics, misassemblies in some of the genome sequences may also be a problem. The *Cannabis* genomes are highly heterozygous, due to a large effective population size and recent hybridization with divergent lineages (Lynch *et al.*, 2017, this issue). The assembly of individuals with high heterozygosity can result in homologous loci with different haplotypes assembled as separate, redundant scaffolds (Vinson *et al.*, 2005; Kajitani *et al.*, 2014). For instance, the *Cannabis* PK assembly

Table 1. Metrics for three available genomic assemblies.

	Genome assemblies						Transcriptome assemblies	
	LA Confidential	NCBI LA Confidential	Chemdog	NCBI Chemdog 91	NCBI PK	NCBI PK scaffold	PK	Finola
Date of Publication	—	2016	—	2016	2011	2011	2011	2011
Size (Mb)	299	595	287	286	**757**	466 (389)*	—	—
Number of contigs	232,373	311,039	175,268	190,122	135,164	**60,029**	—	—
Number of contigs > 1000 bp	95,354	186,080	90,599	90,500	75,416	**57,529**	—	—
Longest sequence	44,496	50,563	73,305	73,305	**556,908**	**556,908**	—	—
Average sequence length	1296	1915	1634	1634	5605	**11,386**	—	—
Contig N50	1603	2648	2249	2249	16,298	**19,021**	—	—
Contig L50	49,511	55,094	33,802	33,770	10,934	**8842**	—	—
Contig N90	626	836	711	710	2628	**5100**	—	—
Contig L90	176,220	221,321	127,329	127,200	52,310	**37,055**	—	—
Unique BUSCO hits bit score > 50	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	0.97	0.99	0.99
Unique BUSCO hits bit score > 100	0.74	0.76	0.74	0.74	**0.76**	0.70	0.99	0.98
Unique BUSCO hits bit score > 200	0.32	**0.33**	**0.33**	**0.33**	**0.33**	0.29	0.89	0.84
Sequencing method	Roche 454	Roche 454	Illumina GA II	Illumina GA II	Illumina GA II; HiSeq	Illumina GA II; HiSeq	Illumina GA II; HiSeq	Illumina GA II; HiSeq
Assembly method	Newbler v.May 2011	Newbler v.May 2011	CLCBio v.5	CLCBio v.5	SOAPdenovo v.1.0.5	SOAPdenovo v.1.0.5	ABySS v1.2.8	ABySS v1.2.8
Genome coverage	50X	50X	300X	300X	130X	130X	—	—
Accession number		GCA_001510005.1		GCA_001509995.1	GCA_000230575.1	AGQN00000000.1	—	—

Note. These assessments allow for the comparison of the available genomes. The best genome assembly, based on each metric, is in bold.
*Size after filtering for redundant contigs.

appears to have been redundantly assembled due to heterozygosity (Lynch *et al.*, 2017). After filtering to remove redundant regions, which were approximately 16.7% of the assembled genome, 389 Mbp remained.

B. Future directions for Cannabis assemblies

Sequences from an inbred individual would improve the current genomic assemblies that may be misassembled due to heterozygosity. Additionally, having the assembly and filtering parameters publicly available will aid in the construction of an improved genome assembly. Assemblies using long-read technologies such as PacBio could also yield improvements. The most dramatic improvements would come from generating sequence-based genetic and physical maps, placing the thousands of contig sequences onto the ten chromosomes.

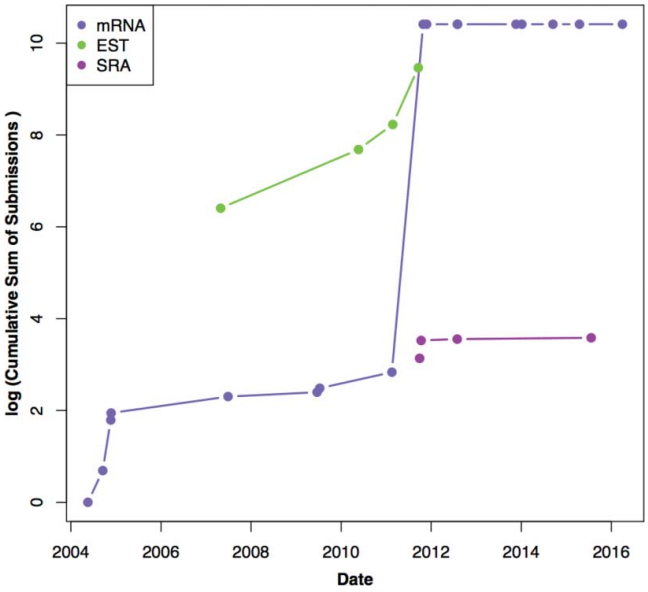

Figure 1. Sum of the submissions of different genomic and genetic tools by date. mRNA data (purple circles) that has been submitted since 2004 is the most widely available. EST, or Expressed Sequence Tag, data (green circles), is a type of mRNA sequence data, which have not been submitted since 2012. SRA data (pink circles) have had a steady submission rate since 2012 and include next-generation sequencing data.

C. Transcriptomes

In addition to the genomic DNA resources available for *Cannabis*, there are a number of transcriptomic, RNA-based resources, which only include RNA being expressed at the time of extraction. The advantages of transcriptomic data are due to their focus on functional, expressed genes, which result in higher coverage of the sequenced sites and/or the ability to sequence more individuals because of the reduction in the number of unique bases to be sequenced. An additional benefit is the ability to quantify both neutral and functional diversity across synonymous and nonsynonymous sites, and variation in expression levels, which indirectly provides details about the evolution of genetic regulatory elements (Elshire *et al.*, 2011).

Compared to other economically important crops, *Cannabis* has little publically available RNA sequence data. For example, at present 106,755 mRNA sequence submissions on the NCBI nucleotide database exist for *Cannabis*, and only 36 transcriptome library submissions are from NCBI's Sequence Read Archive (SRA). Comparatively, *Zea* has 5,136,742 mRNA sequence submissions and 5,050 SRA submissions. Further, the rate of *Cannabis* mRNA sequence accumulation has slowed, with relatively few new submissions since 2011 (Figure 1).

van Bakel *et al.* (2011) assembled two transcriptomes from flowers at different stages (NCBI accessions: SRS266825, SRS266824, SRS266823, SRS266822), roots (NCBI accession: SRS266829) of the marijuana-type cultivar PK, and the flowers (NCBI accession number: SRS266735) of the hemp cultivar Finola (Table 2). The raw reads for each tissue type are available in NCBI, and the two transcriptome assemblies are available from http://genome.ccbr.utoronto.ca/. Although the PK transcriptome assembly includes a variety of tissue types, it is

Table 2. Metrics for the four available mitochondrial genomes and six available chloroplast genomes.

Organelle	Cultivar	Length (bp)	Differences to Carmagnola	%AT	Protein coding genes	NCBI accession	Total genes	Source
Mitochondria	Carmagnola	414,545	—	54	36	KR059940.1	54	White *et al.* (in press)
	Chinese hemp	415,602	257	54	36	KU310670	54	NCBI
	PK	414,737	227	54	—	AGQN00000000.1	—	van Bakel *et al.* (2011)
	Unknown	415,751	408	54	—	—	—	Website
Chloroplast	Carmagnola	153,871	—	63	86	KP274871	127	Vergara *et al.* (2015)
	Dagestani	153,871	16	63	86	KR779995	127	
	Cheongsam	153,848	143	63	86	KR184827	127	Oh *et al.* (2015)
	Yoruba Nigeria	153,854	128	63	86	KR363961	127	
	PK	152,942	22	63	—	AGQN01337109.1	—	van Bakel *et al.* (2011)
	Unknown	153,805	203	63	—	—	—	Website

These assessments allow for the comparison of the available mitochondrial and chloroplast genomes. Differences between the organellar genomes include both SNP and INDEL difference.

comparable to the Finola assembly in terms of its completeness as measured by the proportion of BUSCO genes included in the assembly (Table 1). These BLAST analyses were performed using a BLASTX with default parameters using the same methodology described in the genomic assemblies. Similarly, the number of genes in the single-copy portion of the PK transcriptome assembly is comparable to that of the PK genome assembly. Because of the similarities between the transcriptomic assemblies, with both being quite complete and of high quality, either or both assemblies can be quite useful.

These transcriptome assemblies were used to evaluate variation in the expression of the cannabinoid pathway genes among tissue types in the high- and low-cannabinoid plant varieties PK and Finola, respectively (van Bakel *et al.*, 2011). This study revealed the apparent instances of differential expression between genes in the cannabinoid pathway among the two cultivars. The authors suggested this pattern is most consistent with variation in gene regulation, as they found little evidence for the variation in gene copy number.

D. Future directions for Cannabis transcriptomic data

In the SRA data available for transcriptomic data on *Cannabis*, it is clear that there is a lack of taxonomic representation (Figure 2). Although it is a valuable foundation for further study, van Bakel *et al.*'s (2011) transcriptomic work needs to be replicated to evaluate its generality across the diversity of *Cannabis* varieties, and to make statistically relevant datasets. While there are obvious legal hurdles that hinder further generation of sequence data for *Cannabis*, the recent surge of industrial interest for both medical and agricultural purposes should encourage and fund further studies.

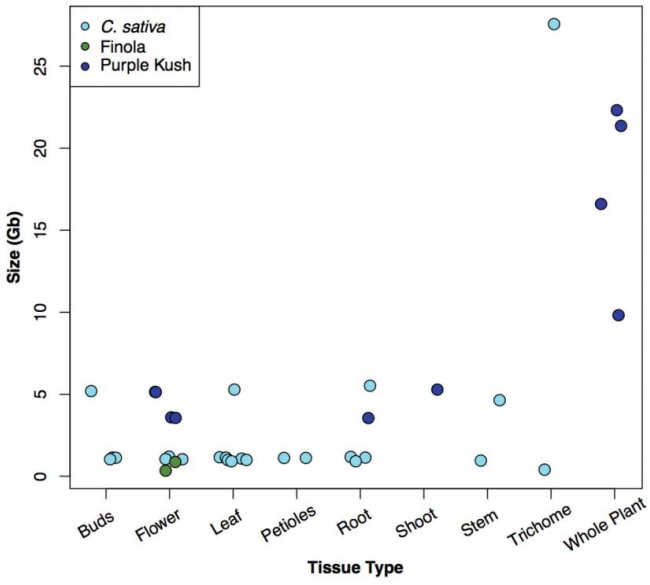

Figure 2. Amount of transcriptomic data available for each tissue type. The amount of transcriptomic data available for a variety of different tissues is very similar, except for whole plant; however, most of it is for only one strain. Blue points identified as *C. sativa* had no further strain information provided in the SRA submission. The tissue types labeled "Whole Plant" were from libraries with unspecified tissue type in the SRA submission.

E. Organellar genomes

The mitochondria and the chloroplasts provide interesting phylogenetic comparisons because of their maternal inheritance in *Cannabis*. Four mitochondrial genome assemblies are accessible, three of which are published on NCBI (Table 2). The four genomes are similar in total length and sequence, and the nucleotides are almost equally distributed (54% AT). Two of the mitochondrial genomes have been fully annotated and belong to hemp cultivars. They have the same number of total genes (54), protein coding genes (36), rRNAs (3), and tRNAs (15) (White *et al.*, 2016). The other two mitochondrial genomes await annotation. One of them is partially assembled, comprises six contigs, belongs to the PK variety, and was sequenced in 2011 (van Bakel *et al.*, 2011). The last mitochondrion is fully assembled and belongs to an unspecified individual (http://www.medicinalgenomics.com/chloroplast-and-mitochondrial-haplotypes-of-cannabis/). We recommend the use of the Chinese hemp or Carmagnola (White *et al.*, 2016) assemblies.

Six chloroplast genome assemblies for *Cannabis* are available, four of which are complete and annotated (Oh *et al.*, 2015; Vergara *et al.*, 2015); one of which is fully assembled but not annotated (van Bakel *et al.*, 2011), and one which is partially assembled and awaits annotation http://www.medicinalgenomics.com/chloroplast-and-mitochondrial-haplotypes-of-cannabis/; Table 2). These chloroplast genomes are very similar to each other in total length, sequence identity, and AT composition (63%). Of the six assemblies, three belong to hemp types, two to marijuana types, and the sixth is an unspecified cultivar. The four annotated chloroplasts have 127 genes, of which 86 are protein coding, 4 are rRNA, and 37 are tRNA (Oh *et al.*, 2015; Vergara *et al.*, 2015). However, Vergara *et al.* (2015) reported 83 coding genes, omitting three, and Oh *et al.* (2015) counted the number of rRNAs from both inverted repeats as unique, reporting each twice. The long single-copy region of the unknown chloroplast is inverted in its orientation with respect to the other published genomes, most likely because its assembly was not oriented canonically, with the origin of replication (*ori*) at the start of the assembly. For chloroplast genomic research, we recommend the use of the

fully assembled and annotated chloroplast genomes published by Oh *et al.* (2015) and Vergara *et al.* (2015).

F. Future directions for the improvement of Cannabis *organellar genomes*

Numerous comparisons such as gene diversity within and between cultivars, as well as within groupings (hemp and marijuana types) remain to be performed using these genomes. Still, both types of organellar genomes show the expected phylogenetic relationships to related taxa in the angiosperms (Oh *et al.*, 2015; Vergara *et al.*, 2015; White *et al.*, 2016). Despite the lack of information in some of the genomes for both the mitochondria and chloroplast, full comparisons between the available data are still possible.

G. GBS data

Genotyping by sequencing (GBS) uses high-throughput, short-read sequences to provide low-cost genotyping with high information content. GBS libraries are produced using restriction enzymes that cut at specific short sequences found many times in the genome. The libraries produced represent ~1% of the whole genome and facilitate genotyping before a reference genome is available (Elshire *et al.*, 2011), and is often more financially accessible than other genotyping options.

Currently there are 325 GBS samples on NCBI's GenBank: 143 samples from the Dalhousie University and 182 samples from the University of Colorado Boulder (Table 3). According to popular classification, the Dalhousie project contains 43 *hemp*, 64 *hybrid*, 7 *indica*, 10 *sativa*, and 19 unclassified sequences (Sawler *et al.*, 2015). The CU Boulder dataset has 10 *hemp*, 65 *hybrids*, 33 *indica*, 17 *sativa*, and 57 unknown samples (Lynch *et al.*, 2017). Of these samples, 133 have been classified genetically (Table 3).

H. WGS data

Whole genome shotgun (WGS) data are created by randomly shearing the entire genome into small fragments, which are then sequenced and reassembled into genomic scaffolds. An advantage of WGS over GBS is that most of the genome is sequenced, so it is easier to compare different datasets, while two different GBS datasets may contain few or no overlapping markers.

Currently, 68 WGS datasets are available for *Cannabis* (Table 3). Of these, 57 are from CU Boulder, divided into 15 *hemp*, 14 *hybrid*, 14 *indica*, and 14 *sativa* samples. While CU Boulder has produced the largest WGS dataset, other sequencing projects are also available (van Bakel *et al.*, 2011; Medicinal Genomics (http://www.medicinalgenomics.com/)). van Bakel and collaborators (2011) sequenced one *indica* and two *hemp* varieties. Medicinal Genomics has eight different strains commonly classified as the following: four *hybrid*, one *indica*, and three *sativa* samples. Sixty-two of these samples have been classified genetically. All of these genomes

Table 3. *Cannabis* genomic datasets.

			Common classification		Genetic classification			
Data type	Provider	Total # samples	Group	# of samples	Group	# of samples	Scientific publication	NCBI accession
GBS	University of Dalhousie	143	Hemp	43	—	—	Sawler *et al.* (2015)	PRJNA285813
			Hybrid	64	—	—		
			Indica	7	—	—		
			Sativa	10	—	—		
			Unknown	19				
	University of Colorado	182	Hemp	10	Hemp	6	Lynch *et al.* (2017)	PRJNA317659
			Hybrid	65	BLDT	45		
			Indica	33	NLDT	82		
			Sativa	17	Unknown	49		
			Unknown	57				
WGS		57	Hemp	15	Hemp	14		PRJNA310948
			Hybrid	14	BLDT	12		
			Indica	14	NLDT	28		
			Sativa	14	Unknown	3		
	Medicinal Genomics	8	Hybrid	4	Hemp	—	—	—
			Indica	1	BLDT	2		
			Sativa	3	NLDT	3		
					Unknown	2		
	van Bakel *et al.* (2011)	3	Hemp	2	Hemp	2	van Bakel *et al.* (2011)	PRJNA73819
			Indica	1	BLDT	1		

Four research groups have sequenced low-coverage full genomes or GBS datasets, three of whom have published their data in peer-reviewed scientific journals and have made their data available through NCBI. Lynch *et al.*'s (2017) full analysis included 340 of the 393 total available samples, and they genetically classified unique 195 cultivars. Even though CU Boulder sequenced a total of 182 genomes, only 169 passed quality control filters and the other 13 were removed from the analysis.

have been sequenced using the Illumina platform except for one, which was sequenced using 454 technology.

I. Future directions for genomic sequencing

The genomic sequences for *Cannabis* can be improved in multiple ways. In the future, sequencing new varieties such as *ruderalis* would greatly improve our understanding of the overall genetic variation in the genus. In addition to WGS data, the two existing GBS datasets cover a large amount of diversity within the *Cannabis* genus, but they were sequenced using different restriction enzymes during their library preparation procedures. This has led to few genomic regions overlapping between both datasets, making their comparison challenging. Therefore, we recommend future sequencing endeavors to include WGS, which will allow for comparisons between future datasets, and to sequence as much variation as possible.

III. Genetic diversity within and among lineages

A. Common vs. genomic classifications

Results generated by genomic studies often contradict the industry's popular classifications of *Cannabis*, which vary between breeders, growers, and dispensaries. Studies concur that *hemp* is a distinct group (van Bakel *et al.*, 2011; Sawler *et al.*, 2015; Lynch *et al.*, 2017), and that two marijuana-type groups may exist (van Bakel *et al.*, 2011; Sawler *et al.*, 2015; Lynch *et al.*, 2017). Lynch *et al.* (2017) used the FLOCK (Duchesne and Turgeon, 2012) model for population structure inferences that established these two marijuana and one hemp groups and suggested a naming convention based on leaf phenotypes shared within them, modified from Clarke and Merlin (2013): narrow-leaf drug type (NLDT), broad-leaf drug type (BLDT), and hemp (Table 3). Using Lynch *et al.* (2017) groupings, 19% of the popularly classified *sativa* clustered within the NLDT grouping, 27% of the *indica* clustered within the BLDT grouping, and 36% and 62% of *hybrids* fell into the BLDT and NLDT categories, respectively. Additionally, cultivars that are popularly reported as 100% *sativa* can be more closely related to those reported to be 100% *indica* (Sawler *et al.*, 2015). The individuals commonly classified as *hemp* fall into this genomic category 86% of the time, suggesting that the industry usually classifies these samples accurately.

If we focus on the specific cultivar names, Sawler *et al.* (2015) found that 35% of their samples were more genetically similar to those with different names than to samples with identical names. Similarly, cultivars with the same name might cluster in different groups, which is the case for *Girl Scout Cookies* and *R4* (Lynch *et al.*, 2017). This is also the case for the cultivar *ACDC*, which is popularly classified as a *hybrid*, but appears to be more closely related to the hemp genetic grouping (Lynch *et al.*, 2017).

B. Hemp

Hemp cultivars appear to be more closely related to each other (van Bakel *et al.*, 2011; Sawler *et al.*, 2015; Lynch *et al.*, 2017). However, it is still under debate whether hemp is more closely related to the colloquial *sativa* or to *indica* groupings. Classic taxonomy proposed hemp as a part of *C. sativa* (Small and Cronquist, 1976), which is supported by recent genomic studies that found the genetic groups hemp and NLDT to be more closely related to each other (Lynch *et al.*, 2017). Still, others found that hemp is more related to the colloquial *indica* grouping (Sawler *et al.*, 2015). All hemp varieties cultivated for either oil or fiber appear to have similar phenotypic characteristics (Anderson, 1980), except for the Asian hemp that seems to be more closely related to *indica* than to *sativa*, or to the BLDT genetic group (Lynch *et al.*, 2017). Thus, the question arises as to whether the hemp strains used by Sawler and collaborators (2015) are mostly of Asian origins, which could explain their close relationship to *indica* plants.

C. Marijuana types

Marijuana-type plants appear to have genetic differences. The STRUCTURE (Raj *et al.*, 2014) and FLOCK analyses by Sawler *et al.* (2015) and Lynch *et al.* (2017) suggest two divisions within the marijuana-type plants. Both groupings contain similar levels of within-group genetic variation (Lynch *et al.*, 2017). However, Lynch and collaborators found that the amount of variation between BLDT and NLDT types is much lower than the variation between either group, and hemp.

D. Comparison between hemp and marijuana types

The genomic studies measured variation between hemp and marijuana types using heterozygosity, but their results differ in the amount of diversity found within these groupings. One study found that hemp contains more heterozygosity than marijuana strains suggesting that hemp varieties are derived from a broader genetic pool (Sawler *et al.*, 2015). Other studies suggest that

hemp contain less heterozygosity due to the possible widespread hybridization of marijuana strains and the stable seed stock required for fiber and seed hemp cultivars (Lynch *et al.*, 2017). However, both studies used a different number of single nucleotide polymorphisms (SNPs) for their analyses.

E. Future directions on Cannabis classification

Thanks to these genomic studies, we know that the colloquial classifications given to the plant are not accurate. The classic taxonomic classifications might also be incorrect, but without proper sample sizes that provide enough genetic variation from all extant *Cannabis* taxa, phylogenetic relationships remain to be established. We therefore recommend that public outreach be used to decrease the knowledge gap between researchers and the general community.

IV. Key biosynthetic pathways

A. Terpenoids

Cannabis produces approximately 100 terpenoids, mostly in the female inflorescences (Potter, 2004, 2009). The terms "terpenoid" and "terpene" are used interchangeably (Gershenzon and Dudareva, 2007) and are a type of secondary metabolite phytochemical excreted by plants that appear to function as defense against herbivory (Langenheim, 1994; McPartland *et al.*, 2000; Sirikantaramas *et al.*, 2005). However, humans have selected for terpenoids by fragrances they produce.

In addition to contributing to the aromatic profile of the plant, these compounds also have biological, biochemical, and medical activities (McPartland and Russo, 2001; ElSohly and Slade, 2005; Mehmedic *et al.*, 2010). At higher concentrations, terpenoids produce psychoactive effects and possibly interact with cannabinoids, altering the effect of both chemicals (McPartland and Russo, 2001; Adams and Taylor, 2010). Recent studies have shown that some terpenoids have sedative effects (Hazekamp and Fischedick, 2012) and show potential as treatments for illnesses such as inflammation (Gil *et al.*, 1989; Lorenzetti *et al.*, 1991), pain (Russo, 2006; Huestis, 2007), depression (Komori *et al.*, 1995), convulsions (Elisabetsky *et al.*, 1995), cancer (De-Oliveira *et al.*, 1997; Vigushin *et al.*, 1998), and fungal infections (Langenheim, 1994).

Without a clear way to define the taxonomy of *Cannabis*, previous research relied on terpenoid variation as a measure of divergence within the genus (Hillig, 2004). However, artificial selection in marijuana and hemp altered the composition of monoterpenoids and other volatile compounds, which are primarily responsible for the aromatic variation among *Cannabis* strains (Ross and ElSohly, 1996; Mediavilla and Steinemann, 1997). Thus, terpenoids likely reflect the history of interbreeding, natural, and artificial selection, rather than purely ancestry.

B. Cannabinoids

Cannabis is perhaps best known for its production of cannabinoids, a class of terpenoids including approximately 25,000 different compounds (Zwenger and Basu, 2008). The plant produces as many as 100 phytocannabinoids, whose compositions vary among cultivars (Brenneisen, 2007; Mehmedic *et al.*, 2010). Cannabinoids are primarily produced in glandular trichomes, which are concentrated in the female flower (Kim and Mahlberg, 1997a, b; Potter, 2004, 2009), with little to no production in other parts of the plant (Dayanandan and Kaufman, 1976; Mahlberg and Kim, 2004). These findings are supported by transcriptomic data establishing that enzymes from the cannabinoid pathway are expressed in the flower (van Bakel *et al.*, 2011), specifically in the glandular trichomes (Marks *et al.*, 2009). Transcriptomic studies have also established that the trichomes in the flower harbor other cannabinoids involved as precursors in the cannabinoid pathway (Gagne *et al.*, 2012).

Cannabis research has been heavily focused on the cannabinoids Δ-9-tetrahydrocannabinolic acid (THCA) and cannabidiolic acid (CBDA) (Russo, 2011). When heated, dried, or stored, these two compounds decarboxylate into the neutral forms of tetrahydrocannabinol (THC) and cannabidiol (CBD), respectively (de Meijer *et al.*, 2003; Russo, 2011). These neutralized forms are able to interact with the brain's endocannabinoid system (McPartland *et al.*, 2006; Russo, 2011) to produce psychoactive effects on invertebrates and vertebrates (McPartland *et al.*, 2006).

Studies on the biosynthetic pathway of THC and CBD synthases have shown similar biochemical properties in both mass and kinetics, suggesting a relationship between the two enzymes (Shoyama *et al.*, 1975; Taura *et al.*, 1995; Taura *et al.*, 1996). It was originally thought that CBDA was converted to THCA (Mechoulam, 1970; Shoyama *et al.*, 1975), but now it is widely accepted that cannabigerolic acid (CBGA) is the precursor molecule to both THCA and CBDA (Taura *et al.*, 1995; Fellermeier *et al.*, 2001). Premature stop codons in the THCA and CBDA synthase sequences or catalytically inefficient synthases lead to a buildup of CBGA (Mandolino *et al.*, 2003; de Meijer and Hammond, 2005).

The gene responsible for the production of THCA is comprised of a single exon with 1638 base pairs (Kojoma *et al.*, 2006; van Bakel *et al.*, 2011). It was previously

thought that the gene encoding THCA and CBDA synthases was a single locus with two codominant alleles (de Meijer *et al.*, 2003; Mandolino and Carboni, 2004), but we will discuss later why this may not be the case. Using this previous model, the differences in the THCA/CBDA ratios were due to either sequence variation in the alleles or to polymorphic expression caused by regulatory elements (de Meijer *et al.*, 2003). Hence, the three distinct chemotypes produced by this Mendelian model suggested the following three different genotypes: one associated with hemp (homozygous for CBDA) and two with marijuana-type plants (homozygous or heterozygous for THCA) (Pacifico *et al.*, 2006; van Bakel *et al.*, 2011). Crosses between chemotypically divergent parents exhibited a 1:2:1 ratio in the F1 progeny supporting this single-locus model (Yotoriyama *et al.*, 1980; de Meijer *et al.*, 2003; Weiblen *et al.*, 2015). Additionally, expression analysis established that the marijuana strain PK expresses THCA synthase and not CBDA synthase (van Bakel *et al.*, 2011), which could be due to the lack of the CBDA allele.

Recent studies suggest that the presence of multiple paralogs is related to the variable production of THCA and CBDA (McKernan *et al.*, 2015; Onofri *et al.*, 2015; Weiblen *et al.*, 2015). Divergence between the coding sequences in these paralogs can lead to functional changes in the expressed proteins (Onofri *et al.*, 2015). The transcription rate of the different paralogs does not appear to be correlated with the production of the cannabinoids; however, regulatory elements might play a role in the production of the chemotype (Onofri *et al.*, 2015). Still, of the nine paralogous copies identified in a hemp (Carmen)–marijuana (Skunk #1) cross, only four of them were actively expressed, suggesting that multiple gene copies, their coding sequences, and expression levels are related to the production of CBDA and THCA (Weiblen *et al.*, 2015). Some of the genomic regions related to the production of these cannabinoids appear to have a great amount of variability, particularly in samples that produce both THC and CBD (Welling *et al.*, 2015). A caveat of these investigations is the use of PCR amplification products, which could lead to ambiguous results given the similarities of these paralogs.

C. Future directions for cannabinoids and other terpenoids

The cannabinoid pathway genes are well studied, although numerous questions remain to be answered. The means of duplication of these paralogs remains unknown, as is the question of how these paralogs influence the phenotype. Although it appears that CBDA synthase is the ancestral form and THCA synthase arose after duplication and divergence (Onofri *et al.*, 2015), much remains to be explored regarding the cannabinoid biochemical pathway. We know that cannabinoid chemotype is paraphyletic—high-CBD lineages are not necessarily related to each other (Sawler *et al.*, 2015; Lynch *et al.*, 2017). Finally, we believe that because of the possible opposing psychoactive effects of CBD and THC (Bhattacharyya *et al.*, 2010; Zuardi *et al.*, 2012; Englund *et al.*, 2013; Niesink and van Laar, 2013), CBDA synthase has historically been selected against in marijuana-type plants. However, due to the newly discovered medicinal properties of CBD, selection for chemotypes high in this cannabinoid has seen a resurgence in recent years. Due to the multiple genes involved in the production of cannabinoids, and perhaps terpenoids as well, we recommend the use of whole genomes for the comparison of these genes instead of isolated techniques such as PCR.

D. Genetic maps

The number of linkage groups in a species should correspond to the number of chromosomes in the organism. However, *Cannabis* has 10 chromosomes and only nine linkage groups were discovered in the first linkage map because this map does not include one of the chromosomes (Weiblen *et al.*, 2015), most likely the sex chromosomes.

For this genetic map, linkage groups were created using two types of markers: 103 amplified fragment length polymorphism (AFLP) loci and 16 microsatellites. Genetic distances in centimorgans (cM), the conventional unit measurement for recombination frequency, were recorded along each chromosome with a 6.10 cM average distance between loci along the whole genome. The total distance covered by all linkage groups in the genome corresponded to 335.7 cM.

Two loci, both on linkage group six, separated by approximately one cM, are associated with the ratio for both THCA and CBDA synthases. These loci corresponding to CBDA and THCA are located at 12.5 and 13.6 cM, respectively, and may play a crucial role in the psychoactive and medicinal effects produced by the plant. One caveat for this finding is that only one plant in the study recombined between the two loci, which might be due to the close proximity between them and to the small sample size. Additionally, the phenotypic ratios produced by the rest of the progeny could be due to the previously proposed single-locus/two-allele model. Previous research has also suggested that the two synthases might be a product of either two loci or one locus with two alleles (van Bakel *et al.*, 2011). Plants that produce high THCA and low CBDA have a nonfunctional CBDA synthase allele under the single-locus model or a

nonfunctional locus under the two loci model (McKernan *et al.*, 2015; Onofri *et al.*, 2015; Weiblen *et al.*, 2015).

One locus on linkage group one may encode for cannabinoid quantity (Weiblen *et al.*, 2015). This linkage group, which is completely separate and segregates independently from the loci from group six, might be associated with the size or density of the trichomes (Weiblen *et al.*, pers. Comm.). The presence of different loci related with the production and quantity of both THCA and CBDA synthases explains the existence of different cannabinoid phenotypes (Weiblen *et al.*, 2015).

E. *Future directions for* Cannabis *genetic maps*

Additional crosses of phenotypically divergent individuals will be the basis for an ultra-high-density genetic map that can be used to place the current genomic assemblies onto the nine pairs of autosomes, and the X and Y sex chromosomes. Genetic variation found in current assemblies will also be associated with numerous physical traits of interest. With no existing quantitative trait locus (QTL) or genome-wide association-mapping studies for the species to date, these maps will be transformative in quantifying associations between existing single-gene markers and traits.

V. The genetics of sex determination in *Cannabis*

The majority of angiosperms are hermaphrodites (having both males and female genotypes), with dioecious species (separate male and female genotypes) representing a small fraction in the reported range of 4%–7% (Lloyd, 1974, 1980; Charlesworth, 1991; Sakamoto *et al.*, 1998; Barrett, 2002; Charlesworth, 2002; Geber *et al.*, 2012; Divashuk *et al.*, 2014; Faux *et al.*, 2014). The evolution of sex to reinforce dioecy generally results in the appearance of morphologically distinct sex chromosomes to ensure 1:1 segregation of males and females (Charlesworth and Charlesworth, 1978; Charlesworth *et al.*, 2005; Ming and Moore, 2007; Ming *et al.*, 2007). Very few dioecious angiosperms have sex chromosomes, and of these only a small number have been studied at a molecular level (Charlesworth, 2002). *Cannabis* is a model system to study the evolution of sex chromosomes, providing a unique snapshot into the early transitions from hermaphroditism to dioecy. Aside from karyotypes (Sakamoto *et al.*, 1998; Divashuk *et al.*, 2014; Razumova *et al.*, 2016), various morphological metrics (Moliterni *et al.*, 2004), and a few sex-specific markers (Mandolino *et al.*, 1999; Faux *et al.*, 2014; Faux *et al.*, 2016), little is known about the genetics underlying sex determination.

Cannabis sativa is a diploid species (2n = 20) that is considered dioecious, with separate male and female genotypes being most common than hermaphrodites; in these sex is a genetically determined trait thought to be governed by distinct X and Y chromosomes (Hirata, 1929; Sakamoto *et al.*, 1998; Peil *et al.*, 2003; Divashuk *et al.*, 2014; Dufaÿ *et al.*, 2014). However, hermaphroditic, monoecious varieties also exist. The latter lack heteromorphic sex chromosomes, yet are interfertile with dioecious plants, indicating a very close relationship between both types (Razumova *et al.*, 2016). Many dioecious types are able to express varying degrees of monoecy in response to chemical, hormonal, and environmental cues. The plasticity in this trait has led to some confusion regarding the genetic basis of sex, and the role of the Y chromosome and autosomes in sex determination in *Cannabis*. Some authors list *Cannabis* as having XY sex chromosomes with an X:0 sex determination system (Faux *et al.*, 2014), which means that the Y chromosome does not contribute to sex determination, and the differences between males and females are produced by the number of X chromosomes present. Others argue that sex is governed by an active Y system and autosomes are not involved (Sakamoto *et al.*, 1998), similar to the sex determination system present in humans. Because *Cannabis* is in the same family as *Humulus lupulus* and this species has an X:0 system (Skof *et al.*, 2012), it is parsimonious to assume that there is an X:0 effect on sex determination in *Cannabis*, but the role of the Y remains to be understood.

Progress is being made in our understanding of the morphological characteristics of the sex chromosomes of *Cannabis*. The sizes of the X and Y chromosomes were evaluated, which revealed that the Y chromosome is larger than the X, with a total haploid genome sizes of 843 and 818 Mbp for males and females, respectively (Sakamoto *et al.*, 1998; van Bakel *et al.*, 2011). The large size of the Y chromosome and the existence of a pseudo-autosomal region that recombines with the X (Peil *et al.*, 2003) support the hypothesis that sex chromosomes are in the early stages of evolution (Yamada, 1943; Sakamoto *et al.*, 1998; Ming *et al.*, 2007). Sex-specific DNA markers that allow for individuals to be tested early in development (Mandolino *et al.*, 1999; Divashuk *et al.*, 2014) revealed polymorphisms in the X chromosome between monecious and dioecious types (Peil *et al.*, 2003), which may explain some of the variabilities in sex expression that has led to the debate regarding the role of the Y in sex determination.

A. *Future directions for* Cannabis *sex-chromosome evolution*

Future work will focus on improving the resolution of the current genetic map and necessary genomic

resources that will aid in identifying the genes underlying sex determination. As previously discussed, this genetic map can place the current genomic assemblies onto the X and Y sex chromosomes in addition to the nine pairs of autosomes. Additional sequencing of both males and hermaphrodites should be performed to approach a better understanding of the molecular basis of sex determination in *Cannabis*.

VI. Conclusions

Cannabis is a crop with agricultural, medicinal, recreational, and industrial applications. The genomic resources available for the study of this plant have allowed us to begin exploring its many scientifically and economically interesting properties. These attributes include the synthetic pathways of its multiple secondary compounds, its nascent sex determination system, and its genomic architecture. There is much room for improvement in these resources, both in the depth and breadth of the datasets as well as in the quality of their analysis.

Further studies will provide invaluable information for evolution history, natural selection, plant ecology, and breeding for improved crops and domestication. Current *Cannabis* genomic research will greatly benefit from the sequencing of a more diverse selection of cultivars and feral varieties to improve our understanding of the phylogenetic relationships between them. Determining the associations between phenotypic and genomic data will improve the decisions made in breeding this crop for specific traits.

The state of research of this important crop remains desperately deficient compared to other modern crops of similar importance, as does the quality of public education and outreach from researchers. As genetic analysis becomes easier and analytical tools become more accessible, collaboration among scientists and industry will also become increasingly important. In this rapidly emerging industry, cooperation is essential to make the best use of our endeavors.

Funding

This research was supported by donations to the University of Colorado Foundation gift fund (13401977-Fin8) and to the Agricultural Genomics Foundation.

Author contributions

DV wrote the organellar genomes and genetic diversity within and among lineages sections, and compiled and lead the review; HB wrote the cannabinoid map section; KMC and CSP wrote the cannabinoid and terpenoid section; KGK wrote the genomic assemblies section; JPM wrote the current available genomes section; SBT wrote the transcriptome section; KHW wrote the sex determination section; NCK conceived and directed the project. All authors contributed to manuscript preparation.

Declaration of interest

DV is the founder and president, and NCK is a board member of the nonprofit Agricultural Genomics Foundation.

References

Adams, T. B., and Taylor, S. V. 2010. *Safety Evaluation of Essential Oils: A Constituent-based Approach*. CRC Press, London, UK.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

Anderson, L. C. 1980. *Leaf Variation Among Cannabis Species from a Controlled Garden*, pp. 61–69. Botanical Museum Leaflets, Harvard University, Boston, MA.

Barrett, S. C. H. 2002. The evolution of plant sexual diversity. *Nat. Rev. Genet.* **3**: 274–284.

Bhattacharyya, S., Morrison, P. D., Fusar-Poli, P., Martin-Santos, R., Borgwardt, S., Winton-Brown, T., Nosarti, C., Mo'Carroll, C., Seal, M., and Allen, P. 2010. Opposite effects of Δ-9-tetrahydrocannabinol and cannabidiol on human brain function and psychopathology. *Neuropsychopharmacology* **35**: 764–774.

Brenneisen, R. 2007. Chemistry and analysis of phytocannabinoids and other Cannabis constituents. In *Marijuana and the Cannabinoids*, pp. 17–49. Humana Press, Totawa.

Charlesworth, B. 1991. The evolution of sex chromosomes. *Science* **251**: 1030–1033.

Charlesworth, D. 2002. Plant sex determination and sex chromosomes. *Heredity* **88**: 94–101.

Charlesworth, B., and Charlesworth, D. 1978. A model for the evolution of dioecy and gynodioecy. *Am Natural* **112**: 975–997.

Charlesworth, D., Charlesworth, B., and Marais, G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**: 118–128.

Clarke, R., and Merlin, M. 2013. *Cannabis: Evolution and Ethnobotany*. University of California Press, Berkeley, CA.

Dayanandan, P., and Kaufman, P. B. 1976. Trichomes of Cannabis sativa L.(Cannabaceae). *Am J Bot* **5**: 578–591.

de Meijer, E. P. M., Bagatta, M., Carboni, A., Crucitti, P., Moliterni, V. M. C., Ranalli, P., and Mandolino, G. 2003. The inheritance of chemical phenotype in Cannabis sativa L. *Genetics* **163**: 335–346.

de Meijer, E. P. M., and Hammond, K. M. 2005. The inheritance of chemical phenotype in Cannabis sativa L.(II): cannabigerol predominant plants. *Euphytica* **145**: 189–198.

De-Oliveira, A. C. A. X., Ribeiro-Pinto, L. F., and Paumgartten, F. J. R. 1997. In vitro inhibition of CYP2B1 monooxygenase by β-myrcene and other monoterpenoid compounds. *Toxicol. Lett.* **92**: 39–46.

Divashuk, M. G., Alexandrov, O. S., Razumova, O. V., Kirov, I. V., and Karlov, G. I. 2014. Molecular cytogenetic characterization of the dioecious Cannabis sativa with an XY chromosome sex determination system. *PloS One* **9**: e85118.

Duchesne, P., and Turgeon, J. 2012. FLOCK provides reliable solutions to the "number of populations" problem. *J. Heredity* **103**: 734–743.

Dufaÿ, M., Champelovier, P., Käfer, J., Henry, J.-P., Mousset, S., and Marais, G. A. B. 2014. An angiosperm-wide analysis of the gynodioecy–dioecy pathway. *Ann. Bot.* **114**: 539–548.

Elisabetsky, E., Marschner, J., and Souza, D. O. 1995. Effects of linalool on glutamatergic system in the rat cerebral cortex. *Neurochem. Res.* **20**: 461–465.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* **6**: e19379.

ElSohly, M. A., and Slade, D. 2005. Chemical constituents of marijuana: the complex mixture of natural cannabinoids. *Life Sci.* **78**: 539–548.

Englund, A., Morrison, P. D., Nottage, J., Hague, D., Kane, F., Bonaccorso, S., Stone, J. M., Reichenberg, A., Brenneisen, R., and Holt, D. 2013. Cannabidiol inhibits THC-elicited paranoid symptoms and hippocampal-dependent memory impairment. *J. Psychopharmacol.* **27**: 19–27.

Faux, A.-M., Berhin, A., Dauguet, N., and Bertin, P. 2014. Sex chromosomes and quantitative sex expression in monoecious hemp (Cannabis sativa L.). *Euphytica* **196**: 183–197.

Faux, A. M., Draye, X., Flamand, M. C., Occre, A., and Bertin, P. 2016. Identification of QTLs for sex expression in dioecious and monoecious hemp (Cannabis sativa L.). *Euphytica* **209**: 357–376.

Fellermeier, M., Eisenreich, W., Bacher, A., and Zenk, M. H. 2001. Biosynthesis of cannabinoids. *Eur. J. Biochem.* **268**: 1596–1604.

Gagne, S. J., Stout, J. M., Liu, E., Boubakir, Z., Clark, S. M., and Page, J. E. 2012. Identification of olivetolic acid cyclase from Cannabis sativa reveals a unique catalytic route to plant polyketides. *Proc. Natl. Acad. Sci.* **109**: 12811–12816.

Geber, M., Dawson, T. E., and Delph, L. 2012. *Gender and Sexual Dimorphism in Flowering Plants*. Springer-Verlag, Berlin.

Gershenzon, J., and Dudareva, N. 2007. The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **3**: 408–414.

Gertsch, J., Leonti, M., Raduner, S., Racz, I., Chen, J.-Z., Xie, X.-Q., Altmann, K.-H., Karsak, M., and Zimmer, A. 2008. Beta-caryophyllene is a dietary cannabinoid. *Proc. Natl. Acad. Sci.* **105**: 9099–9104.

Gil, M. L., Jimenez, J., Ocete, M. A., Zarzuelo, A., and Cabo, M. M. 1989. Comparative study of different essential oils of Bupleurum gibraltaricum Lamarck. *Die Pharmazie* **44**: 284–287.

Gish, W., and States, D. J. 1993. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3**: 266–272.

Hazekamp, A., and Fischedick, J. T. 2012. Cannabis—from cultivar to chemovar. *Drug Test. Anal.* **4**: 660–667.

Hillig, K. W. 2004. A chemotaxonomic analysis of terpenoid variation in Cannabis. *Biochem. Syst. Ecol.* **32**: 875–891.

Hillig, K. W. 2005. Genetic evidence for speciation in Cannabis (Cannabaceae). *Genet. Resources Crop Evol.* **52**: 161–180.

Hillig, K. W., and Mahlberg, P. G. 2004. A chemotaxonomic analysis of cannabinoid variation in Cannabis (Cannabaceae). *Am. J. Bot.* **91**: 966–975.

Hirata, K. 1929. Cytological basis of the sex determination in Cannabis sativa. *Idengaku Zasshi* **4**: 198–201.

Huestis, M. A. 2007. Human cannabinoid pharmacokinetics. *Chem. Biodivers.* **4**: 1770–1804.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., and Maruyama, H. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**: 1384–1395.

Kim, E., and Mahlberg, P. 1997a. Immunochemical localization of tetrahydrocannabinol (THC) in cryofixed glandular trichomes of Cannabis (Cannabaceae). *Am. J. Bot.* **84**: 336–336.

Kim, E.-S., and Mahlberg, P. G. 1997b. Plastid development in disc cells of glandular trichomes of Cannabis (Cannabaceae). *Mol. Cells* **7**: 352–359.

Kojoma, M., Seki, H., Yoshida, S., and Muranaka, T. 2006. DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" Cannabis sativa L. *Forensic Sci. Int.* **159**: 132–140.

Komori, T., Fujiwara, R., Tanida, M., Nomura, J., and Yokoyama, M. M. 1995. Effects of citrus fragrance on immune function and depressive states. *Neuroimmunomodulation* **2**: 174–180.

Langenheim, J. H. 1994. Higher plant terpenoids: a phytocentric overview of their ecological roles. *J. Chem. Ecol.* **20**: 1223–1280.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., and Kristiansen, K. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**: 265–272.

Linnaeus, C. 1753. *Species Plantarum*, vol. 2, p. 1027. Salvius, Stockholm.

Lloyd, D. G. 1974. Theoretical sex ratios of dioecious and gynodioecious angiosperms. *Heredity* **32**: 11–34.

Lloyd, D. G. 1980. Sexual strategies in plants. *New Phytol.* **86**: 69–79.

Lorenzetti, B. B., Souza, G. E. P., Sarti, S. J., Santos Filho, D., and Ferreira, S. H. 1991. Myrcene mimics the peripheral analgesic activity of lemongrass tea. *J. Ethnopharmacol.* **34**: 43–48.

Lynch, R. C., Vergara, D., Tittes, S., White, K., Schwartz, C. J., Gibbs, M. J., Ruthenburg, T. C., Land, D. P., and Kane, N. C. 2017. Genomic and Chemical Diversity in Cannabis. *Crit. Rev. Plant Sci.*, this issue.

Mahlberg, P. G., and Kim, E. S. 2004. Accumulation of cannabinoids in glandular trichomes of Cannabis (Cannabaceae). *J. Ind. Hemp* **9**: 15–36.

Mandolino, G., Bagatta, M., Carboni, A., Ranalli, P., and de Meijer, E. 2003. Qualitative and quantitative aspects of the inheritance of chemical phenotype in Cannabis. *J. Ind. Hemp* **8**: 51–72.

Mandolino, G., and Carboni, A. 2004. Potential of marker-assisted selection in hemp genetic improvement. *Euphytica* **140**: 107–120.

Mandolino, G., Carboni, A., Forapani, S., Faeti, V., and Ranalli, P. 1999. Identification of DNA markers linked to the male sex in dioecious hemp (Cannabis sativa L.). *Theor. Appl. Genet.* **98**: 86–92.

Marks, M. D., Tian, L., Wenger, J. P., Omburo, S. N., Soto-Fuentes, W., He, J., Gang, D. R., Weiblen, G. D., and Dixon, R. A. 2009. Identification of candidate genes affecting delta

(9)-tetrahydrocannabinol biosynthesis in Cannabis sativa. *J. Exp. Bot.* **60**: 3715–3726.

McKernan, K. J., Helbert, Y., Tadigotla, V., McLaughlin, S., Spangler, J., Zhang, L., and Smith, D. 2015. Single molecule sequencing of THCA synthase reveals copy number variation in modern drug-type Cannabis sativa L. bioRxiv: 028654.

McPartland, J. M., Clarke, R. C., and Watson, D. P. 2000. *Hemp Diseases and Pests: Management and Biological Control—An Advanced Treatise.* Cabi Publishing, New York.

McPartland, J. M., Matias, I., Di Marzo, V., and Glass, M. 2006. Evolutionary origins of the endocannabinoid system. *Gene* **370**: 64–74.

McPartland, J. M., and Russo, E. B. 2001. Cannabis and cannabis extracts: greater than the sum of their parts? *J. Cannabis Ther.* **1**: 103–132.

Mechoulam, R. 1970. Marihuana chemistry. *Science* **168**: 1159–1165.

Mediavilla, V., and Steinemann, S. 1997. Essential oil of Cannabis sativa L. strains. *J. Int. Hemp Assoc.* **4**: 80–82.

Mehmedic, Z., Chandra, S., Slade, D., Denham, H., Foster, S., Patel, A. S., Ross, S. A., Khan, I. A., and ElSohly, M. A. 2010. Potency trends of Δ9−THC and other cannabinoids in confiscated cannabis preparations from 1993 to 2008. *J. Forensic Sci.* **55**: 1209–1217.

Ming, R., and Moore, P. H. 2007. Genomics of sex chromosomes. *Curr. Opin. Plant Biol.* **10**: 123–130.

Ming, R., Wang, J., Moore, P. H., and Paterson, A. H. 2007. Sex chromosomes in flowering plants. *Am. J. Bot.* **94**: 141–150.

Moliterni, V. M. C., Cattivelli, L., Ranalli, P., and Mandolino, G. 2004. The sexual differentiation of Cannabis sativa L.: a morphological and molecular study. *Euphytica* **140**: 95–106.

Niesink, R. J. M., and van Laar M. W. 2013. Does cannabidiol protect against adverse psychological effects of THC? *Front. Psychiatry* **4**: 130.

Oh, H., Seo, B., Lee, S., Ahn, D.-H., Jo, E., Park, J.-K., and Min, G.-S. 2015. Two complete chloroplast genome sequences of Cannabis sativa varieties. *Mitochondrial DNA Part A.* **27**: 2835–2837.

Onofri, C., de Meijer, E. P. M., and Mandolino, G. 2015. Sequence heterogeneity of cannabidiolic-and tetrahydrocannabinolic acid-synthase in Cannabis sativa L. and its relationship with chemical phenotype. *Mitochondrial DNA Part A* **27**: 2835–2837.

Pacifico, D., Miselli, F., Micheler, M., Carboni, A., Ranalli, P., and Mandolino, G. 2006. Genetics and Marker-assisted Selection of the Chemotype in Cannabis sativa L. *Mol. Breed.* **17**: 257–268.

Peil, A., Flachowsky, H., Schumann, E., and Weber, W. E. 2003. Sex-linked AFLP markers indicate a pseudoautosomal region in hemp (Cannabis sativa L.). *Theor. Appl. Genet.* **107**: 102–109.

Potter, D. 2004. *Growth and Morphology of Medicinal Cannabis. The Medicinal Uses of Cannabis and Cannabinoids.* Pharmaceutical Press, London.

Potter, D. 2009. *The Propagation, Characterisation and Optimisation of Cannabis Sativa L as a Phytopharmaceutical.* King's College London, London.

Radwan, M. M., Ross, S. A., Slade, D., Ahmed, S. A., Zulfiqar, F., and ElSohly, M. A. 2008. Isolation and characterization of new Cannabis constituents from a high potency variety. *Planta Med.* **74**: 267–272.

Raj, A., Stephens, M., and Pritchard, J. K. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Razumova, O. V., Alexandrov, O. S., Divashuk, M. G., Sukhorada, T. I., and Karlov, G. I. 2016. Molecular cytogenetic analysis of monoecious hemp (Cannabis sativa L.) cultivars reveals its karyotype variations and sex chromosomes constitution. *Protoplasma* **253**: 895–901.

Ross, S. A., and ElSohly M. A. 1996. The volatile oil composition of fresh and air-dried buds of Cannabis sativa. *J. Nat. Prod.* **59**: 49–51.

Russo, E. B. 2006. 11 The Solution to the Medicinal Cannabis Problem. *Ethical Issues Chron. Pain Manage.*: 165.

Russo, E. B. 2011. Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. *Br. J. Pharmacol.* **163**: 1344–1364.

Russo, E. B., and McPartland J. M. 2003. Cannabis is more than simply Δ9-tetrahydrocannabinol. *Psychopharmacology* **165**: 431–432.

Sakamoto, K., Akiyama, Y., Fukui, K., Kamada, H., and Satoh, S. 1998. Characterization; genome sizes and morphology of sex chromosomes in hemp (Cannabis sativa L.). *Cytologia* **63**: 459–464.

Sawler, J., Stout, J. M., Gardner, K. M., Hudson, D., Vidmar, J., Butler, L., Page, J. E., and Myles, S. 2015. The genetic structure of marijuana and hemp. *PloS One* **10**: e0133292.

Shoyama, Y., Yagi, M., Nishioka, I., and Yamauchi, T. 1975. Biosynthesis of cannabinoid acids. *Phytochemistry* **14**: 2189–2192.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Sirikantaramas, S., Taura, F., Tanaka, Y., Ishikawa, Y., Morimoto, S., and Shoyama, Y. 2005. Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. *Plant Cell Physiol.* **46**: 1578–1582.

Skof, S., Cerenak, A., Jakse, J., Bohanec, B., and Javornik, B. 2012. Ploidy and sex expression in monoecious hop (Humulus lupulus). *Botany* **90**: 617–626.

Small, E., and Cronquist, A. 1976. A practical and natural taxonomy for Cannabis. *Taxon* **25**: 405–435.

Swift, W., Wong, A., Li, K. M., Arnold, J. C., and McGregor, I. S. 2013. Analysis of cannabis seizures in NSW, Australia: cannabis potency and cannabinoid profile. *PloS One* **8**: e70052.

Taura, F., Morimoto, S., and Shoyama, Y. 1996. Purification and characterization of cannabidiolic-acid synthase from Cannabis sativa L. Biochemical analysis of a novel enzyme that catalyzes the oxidocyclization of cannabigerolic acid to cannabidiolic acid. *J. Biol. Chem.* **271**: 17411–17416.

Taura, F., Morimoto, S., Shoyama, Y., and Mechoulam, R. 1995. First direct evidence for the mechanism of. DELTA. 1-tetrahydrocannabinolic acid biosynthesis. *J. Am. Chem. Soc.* **117**: 9766–9767.

van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. 2011. The draft genome and transcriptome of Cannabis sativa. *Genome Biol.* **12**: 1241–1250.

Vergara, D., White, K. H., Keepers, K. G., and Kane, N. C. 2015. The complete chloroplast genomes of Cannabis sativa and Humulus lupulus. *Mitochondrial DNA* Part A **27**: 3793–3794.

Vigushin, D. M., Poon, G. K., Boddy, A., English, J., Halbert, G. W., Pagonis, C., Jarman, M., and Coombes, R. C. 1998. Phase I and pharmacokinetic study of D-limonene in patients with advanced cancer. *Cancer Chemother. Pharmacol.* **42**: 111–117.

Vinson, J. P., Jaffe, D. B., O'Neill, K., Karlsson, E. K., Stange-Thomann, N., Anderson, S., Mesirov, J. P., Satoh, N., Satou, Y., and Nusbaum, C. 2005. Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genome Res.* **15**: 1127–1135.

Volkow, N. D., Baler, R. D., Compton, W. M., and Weiss, S. R. B. 2014. Adverse Health Effects of Marijuana Use. *N. Engl. J. Med.* **370**: 2219–2227.

Weiblen, G. D., Wenger, J. P., Craft, K. J., ElSohly, M. A., Mehmedic, Z., Treiber, E. L., and Marks, M. D. 2015. Gene duplication and divergence affecting drug content in Cannabis sativa. *New Phytologist* **208**: 1241–1250.

Welling, M. T., Liu, L., Shapter, T., Raymond, C. A., and King, G. J. 2015. Characterisation of cannabinoid composition in a diverse Cannabis sativa L. germplasm collection. *Euphytica* **208**: 463–475.

White, K. H., Vergara, D., Keepers, K. G., and Kane, N. C. 2016. The complete mitochondrial genome for Cannabis sativa. Mitochondrial DNA Part B **1**: 715–716.

Yamada, I. 1943. The sex chromosome of Cannabis sativa L. *Seiken Ziho* **2**: 64–68.

Yotoriyama, M., Ito, I., Takashima, D., Shoyama, Y., and Nishioka, I. 1980. Plant breeding of cannabis. Determination of cannabinoids by high-pressure liquid chromatography (author's transl). *Yakugaku zasshi: J. Pharmaceut. Soc. Jpn.* **100**: 611.

Zuardi, A. W., Hallak, J. E. C., and Crippa, J. A. S. 2012. Interaction between cannabidiol (CBD) andΔ 9-tetrahydrocannabinol (THC): influence of administration interval and dose ratio between the cannabinoids. *Psychopharmacology* **219**: 247–249.

Zwenger, S., and Basu, C. 2008. Plant terpenoids: applications and future potentials. *Biotechnol. Mol. Biol. Rev.* **3**: 1–7.