XIII Argentine Congress of Bioinformatics and Computational Biology
XIII International Conference of the Iberoamerican Society of Bioinformatics
III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

CHARACTERIZATION OF THE HSP20 SUBFAMILY IN CANNABIS SATIVA

Arredondo S[1*], Lozano A[1*], Calua B[1], Gullino F[1], Chávez U[1], Mantovani L[1], Furque I[1] and Arce, DP[2†]

[1] Instituto Privado "Fray Luis Beltrán" (IFLB), San Nicolás, Buenos Aires, Argentina
[2] Facultad Regional San Nicolás, Universidad Tecnológica Nacional (FRSN-UTN), San Nicolás, Buenos Aires, Argentina
*These authors contributed equally to this work (first authors); † Corresponding author: debora.arce@gmail.com; darce@frsn.utn.edu.ar

INTRODUCTION

HSP20s are 20 kDa heat shock proteins (HSPs) involved in preventing protein aggregation during heat and drought stress responses in numerous plant species. They are conserved across all living organisms, yet their study in Cannabis sativa (Cs) var. hemp has been limited. Cs is one of the earliest domesticated crops, finding applications in various industries such as fiber, food, and oil production. It holds particular significance in medical contexts, especially for pain treatment and addressing refractory epilepsy stemming from neurological disorders. While genotyping and phenotyping of Argentine Cs varieties have only recently commenced, the characterization of CsHSP20 in particular remains ambiguous. Here, we characterized number and location of HSP20s in the latest version of the Cs genome assembly publicly available (Gao et al., 2020).

MATERIALS AND METHODS

Sequence dataset and gene location: HSP20 sequences were retrieved from different databases such as Ensembl plants (https://plants.ensembl.org/index.html), Uniprot (https://www.uniprot.org/), Cannabis GDB (https://gdb.supercann.net/) and NCBI (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/3483/), using "hsp", "hsp20" and "heat shock proteins" as keywords. Redundant sequences were removed after blast alignments and partial results obtained from each consulted platform. Members from the latest genome assembly (cs10 HARVARD OEB (2019) RefSeq:GCF_900626175.2) were kept for further analysis. Genome location was determined using "Genome Data Viewer" from NCBI (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_900626175.2). Tandem repeat genes were considered for gene expression and regulatory analysis by scanning their promoters according to Arce et al. (2018) using "matrix-scan tool" from RSAT plants (https://rsat.eead.csic.es/plants/matrix-scan_form.cgi). Similar methodolgy was conducted to characterize HSP20s from Finola, Chemdog91, CBD18x and Cannatonic2016.

RESULTS & DISCUSSION

This study involves the analysis and characterization of the CsHSP20 subfamily through diverse bioinformatics tools. These tools encompass sequence analysis, chromosomal mapping, gene structure examination, and the utilization of publicly accessible omic data from the Cs genome. Firstly, 40 members of the CsHSP20 subfamily were identified. Then, redundant sequences were removed and located in the latest version genome assembly from NCBI, resulting in a final set of 18 CsHSP20s (Table 1). Noteworthy, the discovery of 8 CsHSP20s situated on chromosome 02, forming a cluster of tandemly arranged genes (Fig. 1). Moreover, the scrutiny of public RNA-seq data revealed that 6 out of these CsHSP20s display upregulation under conditions of drought-induced stress and have HSE in their putative gene promoter regions (Fig. 2).

Table 1. *Characterization of CsHSP20 subfamily in Cannabis sativa (hemp).*

				Physical position on Cs genome	
ID	N	Gene ID from NCBI	Chr	Start position (bp)	End position (bp)
HSP20_1	1	LOC115706297	1	85,171,422	85,175,710
HSP20_2	2	LOC115705042	1	16,088,182	16,088,984
HSP20_3	3	LOC115719406	2	73,361,904	73,362,897
HSP20_4	3	LOC115719530	2	75,207,460	75,208,351
HSP20_5	3	LOC115718620	2	77,290,141	77,290,880
HSP20_6	3	LOC115718619	2	77,302,605	77,303,340
HSP20_7	7	LOC115718621	2	77,314,540	77,315,288
HSP20_8	8	LOC115721022	2	77,322,724	77,323,626
HSP20_9	9	LOC115721203	2	77,345,919	77,346,654
HSP20_10	10	LOC115721265	2	95,193,997	95,194,700
HSP20_11	11	LOC115712164	4	31,444,723	31,445,582
HSP20_12	12	LOC115712088	4	31,448,216	31,449,128
HSP20_13	13	LOC115715573	5	21,975,560	21,980,858
HSP20_14	14	LOC115717061	5	47,035,909	47,041,509
HSP20_15	15	LOC115697017	7	11,645,273	11,645,963
HSP20_16	16	LOC115696221	7	13,386,783	13,387,582
HSP20_17	17	LOC115698828	8	58,220,954	58,221,662
HSP20_18	18	LOC115702178	ND	ctgX47|NW_022060473.1: 25,465 - 26,294	

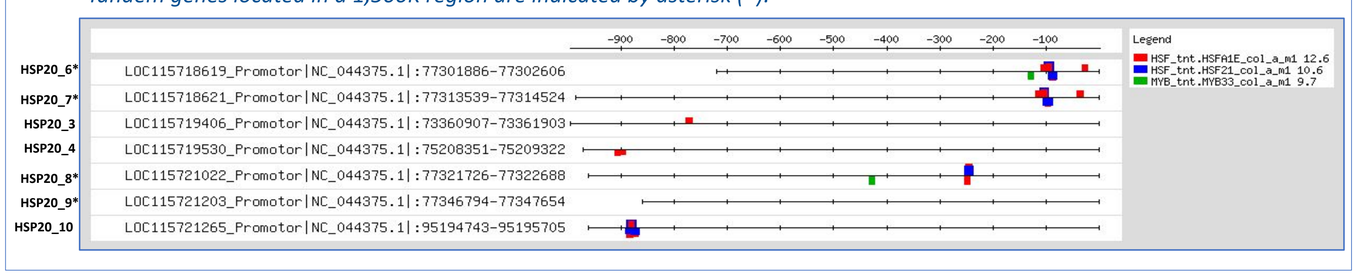
We also identified CsHSP20 members by analyzing Cannabis GDB (Cai et al., 2021). Four varieties were visualized. Hemp (Finola) phenotype shows 59 members, clearly similar to our *in silico* predictions using Ensembl data (40 CsHSP20s in cs10). For phytocannabinoid phenotypes with different concentrations of THC and/or CBD, members varies from 26 to 42 IDs. Further curation by multiple sequence alignment (MSA) is ongoing.

Table 2. *Characterization of CsHSP20 subfamily in five varieties of Cannabis sativa.*

Variety	N CsHSP20	Phenotype	DB
cs10	18	hemp	NCBI (Gao et al. 2020) Annotation release ID: 100.20200513
cs10	40	hemp	Ensembl (van Bakel et al. 2011)
Finola	59	hemp	Cannabis GDB (Cai et al. 2021)
Cannatonic2016	31	phytocannabinoid	Cannabis GDB (Cai et al. 2021)
CBD18x	42	phytocannabinoid	Cannabis GDB (Cai et al. 2021)
Chemdog91	26	phytocannabinoid	Cannabis GDB (Cai et al. 2021)

Figure 1. *Tandemly arranged CsHSP20 genes located in chr02 of cs10 genome (Gao et al., 2020).*

Figure 2. *TFBS in gene promoters of tandemly arranged CsHSP20 from chr02. Location of the HSFA1E (red-squares), HSF21 (blue-squares) and MYB (green-squares) motifs found in the promoter regions (nearly 1K) of eight genes from chr02 in Cs. Tandem genes located in a 1,500K region are indicated by asterisk (*).*



CONCLUSIONS

This study represents the initial instance of characterizing the CsHSP20 subfamily in Cs, shedding light on its involvement in stress responses. The findings underscore the transcriptional activity of these members and the presence of tandemly arranged genes within the Cs genome. The tandemly arranged detected zone could be used for molecular marker designing. Molecular marker technology enables plant breeders to select individual plants based on their marker pattern (genotype) rather than their observable traits (phenotype) and could be used for selection in Cs breeding programs related to phytocannabinoid synthesis.

REFERENCES

Arce, D.P., *et al.* (2018) BMC Genomics, https://doi.org/10.1186/s12864-018-5190-z
Cai, S., *et al.* (2021) Plant Biotechnol J, https://doi.org/10.1111/pbi.13548
Gao, S., *et al.* (2020) Horticulture Research, https://doi.org/10.1038/s41438-020-0295-3
van Bakel, H., *et al.* (2011) Genome Biology, http://genomebiology.com/2011/12/10/R102

Acknowledgements:
Biomedical Engineering GADIB-FRSN-UTN, PID-ICECBSN0008523TC
This study was performed by high school students in their 7th year (computer technicians) IFLB