Characterization of two somaclonal lines of cucumber (Cucumis sativus L.) through a transcriptomic approach

Estelle Bystrzycki

AGROCAMPUS OUEST

☑ **CFR Angers**

☐ **CFR Rennes**

Année universitaire : 2018-2019 Spécialité : Horticulture Spécialisation (et option éventuelle) : Science et Ingénierie du Végétal (SIV), option Semences et plants : recherche et développement, production, commercialisation (SEPRO)	**Mémoire de fin d'études** ☑ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage ☐ d'un autre établissement (étudiant arrivé en M2)

Characterization of two somaclonal lines of cucumber (*Cucumis sativus* L.) through a transcriptomic approach

Par : Estelle BYSTRZYCKI

Soutenu à Angers le 17 septembre 2019

Devant le jury composé de :

Président : Olivier LEPRINCE

Maître de stage : Magdalena PAWEŁKOWICZ

Enseignant référent : Agnès GRAPIN

Autres membres du jury

Jérôme VERDIER

ACKNOWLEDGMENT

I want to thank professor Andrzej Przybyła for all he taught me as an Erasmus student and for being so kind and helpful in my research of an internship in Warsaw University of Life Sciences. Without him, I would not be where I am today. Thank you also to professor Wojciech Pląder who accepted and welcomed me in the Department of Plant Genetics, Breeding and Biotechnology within the Faculty of Horticulture, Biotechnology and Landscape Architecture as a trainee for my Master 1 internship and for my final year internship. Thank you to my amazing supervisor, doctor Magdalena Pawełkowicz who warmly welcomed me both times I came for my internships. She was always there to guide me when I was lost, to push me when I thought I could not go further and to help me when I needed it the most, especially during the hard times of writing this master's thesis. Thank you to the PhD student, Agnieszka Skarzyńska, who had to cope with me almost every day during these two internships and who became a friend. She was as helpful as Magda. Thank you for showing me all the laboratory techniques that I did not know, for helping me to understand the unexpected results that could sometimes drive me crazy, and for talking about anything when the mood was down, for all this laughs that I will never forget. It belongs to these internships as well as all the work we accomplished all together. Thank you to Agnieszka and Magda for believing in me. Thank you to Maciek, Tomek and Krzysiek, the students who were writing their thesis as the same time as me. They helped me in many ways for my experiments and the analysis of the data all along both internships. Exchanging with students from a slightly different field of study made me learn a lot and helped me extend my knowledge. I wish good luck to my friends, Maciek with the pursuit of your studies, Tomek with you PhD in Norway. Thank you to my colleague and friend who was also in internship this year, Joe, for distracting me from my work sometimes and being such a good neighbor. Thank you to all the professors, doctors and PhD students of the department for such a great atmosphere, for sharing your material when ours was broken, or not working. A special thank-you to doctor Ewa Siedlecka who taught me so much in Erasmus and who was always smiling, in a very good mood and so kind during my internships. A special thank you also to doctor Marek Koter for his unwavering good mood and kindness at any time (and thank you for this exceptional laugh that always brought our smiles back on our faces!).

I want to thank, from my school, the National Institute of Horticulture and Landscape Architecture, Olivier Leprince, the supervisor of the option I chose. Thank you for teaching me so many things about the seeds this year but also for 3 years already. He is one of the professors who made me develop my interest in plant physiology and so, who unconsciously guided me to where I am today. Thank you for reacting so fast to any problem we could have during our year of Master 2, and even until the end of our internship. He was always there to fix the problems as well as he could. Thank you to my tutor, Agnès Grapin, for all the precious advices she gave me to write my master's thesis. I took everything I could into account, and I hope she will be satisfied about my work.

However, the greatest thank-you I will give will be for my family and my boyfriend. My mother Aude, my father René, my grandmother Eveline, my godmother Anny and my boyfriend Wojtek were the best supporters I have ever had. Thank you for these hours spent on the phone when I was away, for the time spent together in France or in Poland. Thank you for believing in me. The time I lost during these moments was not really lost but provided me the strength to go on and end my master's thesis. I am sorry if sometimes I was in a bad mood, I was stressed. But they coped with me without complaining. So, THANK YOU!

LIST OF ABBREVIATIONS

ABA: ABscisic Acid
ABRE: ABA-Responsive Element
AP2: APetala 2
BAP: 6-BenzylAminoPurine
bHLH: basic Helix-Loop-Helix
BLAST: Basic Local Alignment Search Tool
bZIP: basic leucine ZIPper
Cas9: CRISPR associated protein 9
cDNA: complementary DNA
CDS: Cytokinin-Dependent Suspension
CES: Cytokinin-dependent Embryogenic Suspension
CMV: Cucumber Mosaic Virus
CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
DEG: Differentially Expressed Gene
DGE: Digital Gene Expression
DLR: Direct Leaf Regeneration
DNA: Deoxyribonucleic Acid
DNase: DeoxyriboNuclease
dNTP: deoxyribose Nucleoside TriPhosphates
DOF: DNA-binding One zinc Finger
EMS: Ethyl MethaneSulfonate
ERF: Ethylene Response Factor
FAO: Food and Agriculture Organization of the United Nations
FDR: False Discovery Rate
gRNA: guide RNA
GWAS: Genome-Wide Association Study
HCL: Hierarchical Clustering
HD-Zip: HomeoDomain-leucine Zipper
IPA: 3-IndolePropionic Acid
jmjC: jumonji C
KAN: KANADI
LCR: Leaf Callus Regeneration
lincRNA: long intergenic non-coding RNA
MADS: *MCM1* (from *Saccharomyces cerevisiae*), *AGAMOUS* (from *Arabidopsis thaliana*), *DEFICIENS* (from *Antirrhinum majus*), *SRF* (from *Homo sapiens*)
miRNA: microRNA
MQ: Milli-Q®
mRNA: messenger RNA
MS: Musharige and Skoog
MYB: MYeloBlastosis
MYC: MYeloCytomatosis
NAC: Nascent polypeptide-Associated Complex
NCBI: National Center for Biotechnology Information
NGS: Next Generation Sequencing
PCR: Polymerase Chain Reaction
PLATZ: PLant AT-rich sequence and Zinc-binding proteins
PPI: Protein-Protein Interaction
qPCR: quantitative PCR
RAPD : Random Amplified Polymorphic DNA
RNA: RiboNucleic Acid
RNA-seq: RNA-sequencing
ROS: Reactive Oxygen Species
rRNA: ribosomal RNA
snoRNA: small nucleolar RNA
SSR: Simple Sequence Repeat
STRING: Search Tool for the Retrieval of Interacting Genes/Proteins
TAE: Tris base, Acetic acid and Ethylenediaminetetraacetic acid (EDTA)
T-DNA: transfer DNA
TF: Transcription Factor
TIP41: TAP42 Interacting Protein of 41 kDa
TPM: Transcripts Per Million
TPS21: TerPene Synthase 21
UBIep: UBIquitin extension protein
UV: UltraViolet
2,4-D: 2,4-Dichlorophenoxyacetic acid

LIST OF FIGURES

Figure 1: Scheme presenting how to obtain the somaclonal lines S2 and S3 from the cucumber B10 line

Figure 2: Photographs showing the comparison between the somaclonal lines S2 and S3 and the wild type B10 from different angles and at different stages of growth

Figure 3: Venn diagram presenting the number of DEGs in S2 and S3 line and the number of common DEGs between the two somaclonal lines

Figure 4a: RNA-seq data, expressed in TPM and represented by the black line, in comparison with the qPCR results, expressed as the relative normalized expression of the S2 DEGs and represented by the grey bars

Figure 4b: RNA-seq data, expressed in TPM and represented by the black line, in comparison with the qPCR results, expressed as the relative normalized expression of the S3 DEGs and represented by the grey bars

Figure 5: Protein network showing the connections between the DEGs of **a)** S2 line and **b)** S3 line.

Figure 6: Abundance of each motif found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line

Figure 7: Distribution of the motifs' length found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line

Figure 8: Distribution of the organisms to which the motifs found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line were attributed

Figure 9: Distribution of the functions of the motifs found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line

Figure 10a: Chromosome maps of all the cucumber chromosomes indicating the position of the DEGs of S2 line on the chromosome and more precisely on the contig. The table gives the number of contigs and DEGs mapped on every chromosome

Figure 10b: Chromosome maps of all the cucumber chromosomes indicating the position of the DEGs of S3 line on the chromosome and more precisely on the contig. The table gives the number of contigs and DEGs mapped on every chromosome

Figure 11: Heatmaps of **a)** S2 and **b)** S3 DEGs. The HCL analysis results are shown on the left of each heatmap

LIST OF ANNEXES

Annex I: Thermocycler programs for cDNA synthesis, PCR and qPCR

Annex II: List of chosen genes to verify the RNA-seq data by qPCR in S2 and S3 lines

TABLE OF CONTENTS

I. INTRODUCTION……..p.1

I.1. Cucumber's economic value in Poland……...p.1

I.2. Next Generation Sequencing (NGS) methods in breeding……p.1

I.3. Source of plant material for breeding…….p.3

I.4. Somaclonal variation……...p.4

I.5. Somaclonal variation in cucumber…….p.5

I.6. Aim of the study……p.7

II. MATERIALS AND METHODS……..p.8

II.1. Plant material……..p.8

II.1.1. *In vitro* culture……..p.8

II.1.2. Greenhouse growth……..p.8

II.2. RNA isolation and cDNA synthesis……...p.8

II.3. Bioinformatic network modeling of DEGs……...p.9

II.4. Verification of RNA-seq data by qPCR…….p.9

II.5. qPCR data treatment……...p.10

II.6. Heatmap construction……..p.11

II.7. Bioinformatics analysis of the patterns in the gene's promoters……...p.11

II.8. DEGs' location on chromosomes…….p.11

III. RESULTS……...p.12

III.1. Phenotypes of the somaclonal lines……p.12

III.2. Statistics on differentially expressed genes……...p.12

III.3. Verification of RNA-seq results by qPCR…….p.12

III.4. Analysis of protein interaction through bioinformatics tools……..p.16

III.5. Bioinformatics analysis of promoter regions of DEGs in somaclonal lines.p.16

III.6. Chromosomal location of DEGs…….p.23

III.7. Heatmap of DEGs……...p.23

IV. DISCUSSION…….p.27

IV.1. Phenotypical analysis of the somaclonal lines of cucumber…….p.27

IV.2. Statistics on differentially expressed genes……p.27

IV.3. Verification of RNA-seq results by qPCR line……..p.27

IV.4. Analysis of protein interaction through bioinformatics tools……..p.28

IV.5. Bioinformatics analysis of promoter regions of DEGs in somaclonal lines.p.30

IV.6. Chromosomal location of DEGs……p.33

IV.7. Heatmap of DEGs……p.34

V. CONCLUSION……p.35

BIBLIOGRAPHY……p.36

ONLINE RESSOURCES……p.45

ANNEXES……p.46

I. INTRODUCTION

The cucumber is a plant from the Cucurbitaceae family, that contains around 750 species. It belongs to the *Cucumis* genus and the latin name of the species is *Cucumis sativus* L. (Malepszy and Niemirowicz-Szczytt, 1991).

It is native to southern Asia, particularly India where it has been cultivated for over 3000 years. While the presence of cucumber extended to European countries such as Greece and Italy, it was introduced at the same time in China. It came into other European countries through the development of the roman empire and appeared for instance in France in the ninth century (INFOAGRO SYSTEMS, SL, 2014).

I.1. Cucumber's economic value in Poland

Cucumber is a very important crop since it has a high rate of consumption. Thus, to respond to the high demand, countries need to have a high offer. The production share of cucumbers is 85,9% for the Asian continent that is therefore the first producing continent, but Europe stands at the second place with an honorable production share of 8,9% (FAO, 2019). In 2011, China was the biggest world producer of cucumber with 47,31 million tons representing 73,16% of the world production. Besides, the Russian Federation, Ukraine and Poland are respectively the fourth, fifth and eleventh biggest producers of cucumber in the world, and first, second and fourth biggest producers in Europe (INFOAGRO SYSTEMS, SL, 2014). Moreover, Poland produced in average, on the last 15 years, more than 161 thousand tons of cucumbers while other European countries like France or Italy produced between 43 thousand tons and less than 161 thousand tons of cucumber, and some other countries even less (FAO, 2019). Therefore, cucumber has a high importance in European countries and especially in Eastern Europe.

Indeed, Poland had an increasing production of cucumber from 2011 to 2017, reaching around 545000 tons in 2017 out of the 83,75 million tons produced in the whole world (INFOAGRO SYSTEMS, SL, 2014; FAO, 2019). The increase of production started in 2010 and is going along with a high increase of the yield (FAO, 2019).

This might be explained by the interest of Polish farmers and people in this plant production. The cucumber is indeed one of the ten most produced commodities in the category of vegetables primary (FAO, 2019) and so it is important for farmers to get a better yield.

I.2. Next Generation Sequencing (NGS) methods in breeding

The NGS techniques enable the sequencing of whole genomes and transcriptomes, and thus the discovery of new genes and their position. Moreover, the sequencing is based on a whole population and not only single individuals. These techniques are able to perform a massive sequencing work producing big data (Pawełkowicz *et al.*, 2016).

Nowadays, the breeders rely more and more on genetic data, where the NGS methods can help by making available a high number of molecular markers for instance. For the marker-assisted selection it is essential to connect the phenotype to the responsible gene(s), and NGS

techniques facilitate the process by providing the genome sequence of the target plant with its structural and functional annotation. In this way, it is easier to establish a correlation between a phenotypical trait and the genes implied in its expression. A very precise gene mapping is then very important (Pawełkowicz *et al.*, 2016). In 2011, the first cucumber gene map was created. SSR markers were used to create a linkage map divided in 7 groups, each corresponding to one of the cucumber chromosomes. The map possesses 248 SSR markers and 7 important traits for horticulture were identified (Miao *et al.*, 2011). A new genetic map was created later, in 2015 (Xu *et al.* 2015). The easier creation of new gene maps enabled a marker-assisted selection of 90% precision of the yellow fruit flesh trait encoded by the *yfl* gene mapped on the 7^{th} chromosome (Lu *et al.*, 2015). However, it is necessary to build genetic maps specific to the variety or line studied, because of the potential chromosome rearrangements (Yang *et al.*, 2012). Among all the cucumber traits that have a breeding importance, several were already studied and deciphered thanks to NGS methods. Moreover, cucumber, as many other plants, is subject to various diseases often provoked by pathogens. However, some genes specialized in the defense against pathogens exist. These genes are called R-genes (Resistance genes). Thanks to NGS, through the use of the already sequenced genome, 70 homolog genes were found in the cucumber and the scientists achieved to delimit 67 of them (Yang *et al.*, 2013), which is a great advance in the fight against pathogens. Nevertheless, the most important traits of the cucumber plant concern the appearance and taste of the fruit (Pawełkowicz *et al.*, 2016). The ancient wild cucumber varieties were bitter, and Man, through domestication, selected fruits that are less and less bitter. Cucurbitacins, a type of triterpenoids, are components that gives a bitter taste to the cucumber flesh. Nine genes implied in the biosynthesis pathway of this components were identified and 2 transcription factors (TF) *Bl* (bitter leaf) and *Bt* (bitter fruit) regulating this pathway were discovered through Genome-wide association study (GWAS). Four catalytic steps were also enlightened (Shang *et al.*, 2014). These examples show that NGS techniques help to find more easily new material for breeding programs regardless of the target trait or its complexity.

With NGS, it is not only possible to build very accurate genetic map, but also to study gene expression (Pawełkowicz *et al.*, 2016). When gene expression is known, it helps the breeder to understand the underlying mechanisms under the expression of a complex trait and thus to find new target genes that can modify this trait (Perez-de-Castro *et al.*, 2012). The most used technology to study gene expression was the microarray technology but RNA-seq is now more widely used as it overcomes the principal limits of the previous method (Perez-de-Castro *et al.*, 2012). NGS technique was used to study lots of different cucumber important breeding traits (Pawełkowicz *et al.*, 2016). Among these traits, parthenocarpy is a characteristic that determines yield and quality. The study highlights that the more active genes are related to cell division. Moreover, it was discovered that the set of a parthenocarpic fruit is a process consuming a lot of sugar through an enhanced carbohydrate degradation (Li *et al.*, 2014). As one of the most complex traits of cucumber, the sex determination of flower is already well studied and the identified genes implied in it were involved in biogenesis, transport and organization of cellular component, macromolecular and cellular biosynthesis, localization, establishment of localization, translation and other processes. The conclusion of this study was that genes involved in hormone signaling pathways and some TF were the most important concepts (Wu *et al.* 2010). Furthermore, in 2010, Guo *et al.* sequenced the transcriptome of cucumber flower buds in two almost isogenic cucumber lines. In another study, 310 differentially expressed genes (DEGs) were identified among hermaphrodite, female and male

cucumber isogenic lines. These DEGs were implied in already known processes (hormone processes and signaling, lipid and sugar metabolism) but new processes were also discovered like cell wall, membrane and cytoskeleton modifications (Pawełkowicz *et al.*, 2019). More than the traits, the full reaction of a plant under certain conditions can be studied. The whole transcriptome of cucumber was sequenced under N deficiency. Among the 23000 studied transcripts, 364 DEGs were identified, of which 64 were related to signaling (Zhao *et al.*, 2015). The root formation in response to melatonin under salt stress was also studied with RNA-seq, resulting in 121 genes upregulated and 196 genes downregulated. Among these DEGs, peroxidase-related genes, several TF families, and genes related to cell-wall formation, carbohydrate metabolic processes, oxidation/reduction processes and catalytic activity were identified (Zhang *et al.*, 2014). The surface of cucumber can be covered with some sort of spines called trichomes. Their presence affects the quality of the fruit since consumers sometimes prefer smooth fruits. A study using digital gene expression (DGE) analysis showed that this trait might be regulated by processes involving meristem genes and polarity regulators (Chen *et al.*, 2014). Fruit development was studied through the development of a cucumber transcriptome, thanks to 454-pyrosequencing after RNA isolation and cDNA synthesis. The most highly expressed genes were compared to *Arabidopsis* genes and the homologs were related to growth, lipid, latex and defense (Ando and Grumet, 2010). RNA-seq methods can also be used to sequence miRNA that are involved in plant development and stress response (Pawełkowicz *et al.*, 2016). Several studies were carried out to find which families of miRNA can be found in cucumber (Martínez *et al.*, 2011), also depending on the organ like leaves and roots (Mao *et al.*, 2012), to understand their role in the fruit formation (Ye *et al.*, 2015), or to determine if grafting has an influence on the miRNA (Li *et al.*, 2013). The technique extends to lincRNAs (long intergenic non-coding RNAs), and in the study it was found that they were involved in response to stimuli, multi-organism processes, reproduction, reproductive processes, and growth (Hao *et al.*, 2015).

I.3. Source of plant material for breeding

Despite the vast genetic diversity of some plant species, breeders are still looking for new sources of plant materials. Moreover, in cucumber, it is very important for breeding programs since its genetic diversity is very narrow (Pawełkowicz *et al.*, 2016). To reach this aim, several methods are available.

The chemical mutagenesis is a rather old method that produces random mutations under the application of chemical products. It was used on cucumber with ethyl methanesulfonate (EMS) and several mutants classified in 6 categories were obtained: short-fruit, long-fruit, small-flower, big-flower, opposite-tendril and clustered-leaf. The short fruit and clustered leaf mutants were thought to be controlled by a single recessive gene while the long fruit mutant was thought to be a homozygous (Wang *et al.*, 2014). Another study, based on the same mutagen agent, obtained dwarf mutants (Shah *et al.*, 2015). Chemical mutagenesis was also used to create mutants resistant to the cucumber mosaic virus (CMV) (Aimin *et al.*, 2004).

The somaclonal variation can also be a source of new genetic material for breeders, even if it was first considered as an undesirable consequence in plant propagation by *in vitro* culture. However, like chemical mutagenesis, it is a rather old method that gives random mutations. It is consequently a very inaccurate method when a precise feature is targeted, as well as the latter

technique. For instance, in an old study, somaclonal variation was proposed as a mean to overcome the post-fertilization barriers in interspecific crosses of *Cucumis* (Custers and Bergervoet, 1984).

Transformation, *i.e.* creation of genetically modified or transgenic plants, is a more recent method that gives less random changes. Indeed, a portion of desired DNA is introduced in the plant, but it is impossible to know where it will be inserted and in how many copies. Moreover, this insertion can create a disorder in other genes expression depending on the place of insertion. Several ways can be used to create transgenic plants but the most widely used nowadays is the insertion of a T-DNA made of a genetic construction containing the gene of interest through an *Agrobacterium* infection (Wang *et al.*, 2015). Cucumbers were transformed in this way for various purposes. In the study of Chen *et al.* (2019), an enhanced photosynthesis and biomass yield were obtained. In another study, researchers tried to enhance the chilling tolerance of cucumber by inserting a dehydrin gene in its genome, but the results were not conclusive in the phytotron conditions and they are still working on this topic (Mróz *et al.*, 2015). Nevertheless, the insertion of the thaumatin II gene from *Thaumatococcus daniellii* in the cucumber genome resulted in the production of sweeter fruits (Szwacka *et al.*, 2002).

The CRISPR/Cas9 technology is the newest of all cited before. It enables directed mutations or insertion of genetic material. It is a construct made of an endonuclease and a gRNA that guides the construct to the desired place in the genome. The endonuclease cuts the DNA at this place and it is then the DNA repairing mechanism which can make an error and thus create a mutation; or a gene was inserted with the construct and the DNA repairing mechanism uses this gene to repair the DNA where it was cut. In 2016, a broad virus resistance was established in cucumber resulting in a non-transgenic virus-resistant plant (Chandrasekaran *et al.*, 2016). Another study resulted in the generation of a transgene-free gynoecious cucumber plant that will be useful for heterosis breeding (Hu *et al.*, 2017).

However, the two latter methods, using transformation technologies, cannot be done in an entire organism and thus, the modified cell(s) need to be regenerated *in vitro* after the transformation. Besides, as described before, during *in vitro* culture, there can be some undesirable mutations called somaclonal variation. This interesting phenomenon was already the target of numerous projects, with the aim to limit it as much as possible so the regenerated plants are true-to-type.

I.4. Somaclonal variation

"Somaclonal variation is the variation seen in plants that have been produced by plant tissue culture" (Bhatia and Sharma., 2015). These variations can be the result of genetic or epigenetic changes compared to the basic cultivar (Guo *et al.*, 2007; Acquaah, 2012). Thus, they can be transient (only visible in the plant that was cultivated *in vitro*) or heritable (the character stays through generations). Every plant cultivated *in vitro* can be subject to somaclonal variation, but some genotypes are more likely to show some changes in such conditions. For instance, polyploid plants are more affected by somaclonal variation than diploid plants (Acquaah, 2012).

Moreover, different factors are known to influence the appearance of somaclonal variations. It is well known that a plant is more likely to express somaclonal variation if it stays

in *in vitro* culture for a long time or if the number of subcultures is too high especially if it is in a callus state (Hao and Deng, 2002; Jevremović *et al.*, 2012; Khan *et al*., 2011). Also, the composition of the medium can have an impact on these variations. Thus, the presence of some auxins, cytokinins or other plant growth regulators and the replacement of photosynthesis products by sugar in the medium could enhance their appearance (Acquaah, 2012; Cassels *et al.*, 2001; Smulders and de Klerk, 2011). Other factors like the wound made or the exposure to the sterilants when the explant is prepared, the fact that the cultivated tissue is incomplete, the disturbed balance between the high humidity and transpiration and the lighting conditions may affect the variation as well (Cassells and Curry, 2001). The tissue source influences the frequence and the nature of the somaclonal variations too (Krishna *et al.*, 2016).

Somaclonal variation can be used in breeding programs to discover how wide can be the variability of a species and is now commonly used in breeding practices for ornamental plants (Krishna *et al.*, 2016). Moreover, the time of selection can be substantially reduced compared to the classical selection programs in the field and can complement the latter (Jain, 2001). A lot of different characters such as biotic and abiotic stress resistance or tolerance (drought, salinity, pH, diseases, etc.) can be selected (Yusnita *et al.*, 2005; Krishna *et al.*, 2016) and a lot of cultivars among different species were developed, particularly in banana, potato and strawberry but also in tomato, pineapple or *Geranium* subspecies (Krishna *et al.*, 2016).

I.5. Somaclonal variation in cucumber

Regenerated cucumber plants from *in vitro* culture can display morphological changes in the color or shape of the leaves, the growth habits, the time of flowering or even the size of the plant organs. The latter can be explained by a change in the ploidy level (Custers *et al.*, 1990). These changes usually have a low impact on the cucumber's quality. However, some variations such as the lower production of seeds can have a definitely negative impact on the production but also on breeding programs.

The selfed progeny of these regenerated plants displays a lower germination rate and only very few of the plants show some somaclonal variation with the same characters as quoted previously for the regenerated plants. Nevertheless, the tetraploidy character was preserved in the selfed progeny (Custers *et al.*, 1990).

Indeed, regeneration methods from a callus culture, a suspension or protoplats imply a rather high rate of abnormal regenerants and somaclonal variation (Malepszy, 1988; Malepszy and Nadolska-Orczyk, 1989)

In 1995, Burza and Malepszy carried out experiments on a new regeneration method for cucumber. This method consisted on regenerating a cucumber plant directly from a leaf explant, without going through a callus stage. The aim was to try this new method and determine if it was more reliable to regenerate a cucumber plant. That is how they found out that, with this method, they might obtain only two new phenotypes in the R1 generation. However, these phenotypes are not stable because absent in the R2 generation.

All the previous studies were aimed at trying to reduce somaclonal variation as much as possible to obtain true-to-type plants. *In vitro* culture was mostly used to regenerate or multiply plants to obtain clones of the cultivated cultivar, and to reach this goal with a good yield and not much losses, somaclonal variation should be as low as possible. However, a few years later,

scientists start to get interested in somaclonal variation on itself and studied it to determine which factors can influence it.

Concerning the study of somaclonal variation in cucumber, the first experiments aimed at studying the relationship between the regeneration system and the genetic variability by measuring the rate of somaclonal variation for each method of regeneration tested (Pląder *et al.*, 1998). A few years later, this topic was dealt with in depth by testing other regeneration methods, different times of culture and modifications of NH_4^+/NO_3^- ratio in the MS medium (Ładyżyński *et al.*, 2002). In total, in these two papers, eight different regeneration methods were tested to result in the same conclusion: the rate and nature of somaclonal variations strongly depends on the regeneration system and the other parameters like the time of culture and the composition of the chosen medium (Pląder *et al.*, 1998; Ładyżyński *et al.*, 2002). Moreover, these studies lead to the discovery and the further studies of somaclonal lines, starting with a previously described variant of msc (mosaic cucumber phenotype) characteristics (Malepszy *et al.*, 1996).

Finally, three somaclonal lines were obtained from the same cucumber cultivar called Borszczagowski: the line named S1 resulting from direct leaf regeneration (DLR) that was the subject of another work, the line named S2 resulting from leaf callus regeneration (LCR) and the line named S3 resulting from cytokinin-dependent embryogenic suspension (CES) using shoot apical buds that were first put on an MS medium for three weeks (**Fig. 1**). The two latter will be studied in this work.

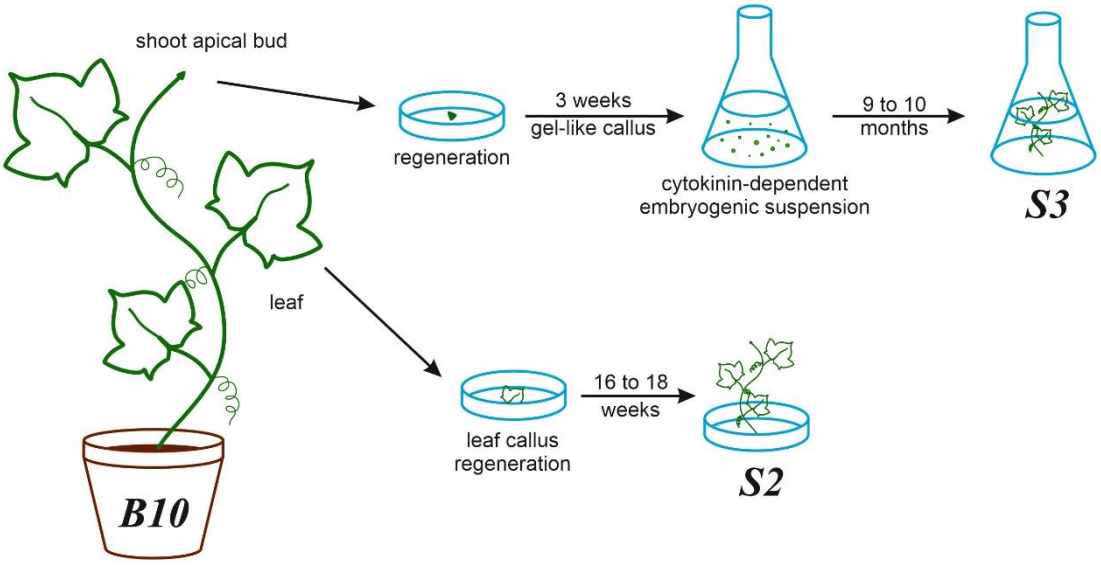


Figure 1: Scheme presenting how to obtain the somaclonal lines S2 and S3 from the cucumber B10 line

More recently, the assembly of these somaclones was realized (Skarzyńska *et al.*, 2017) as well as the characterization of S1 somaclonal line through a transcriptomic approach and thanks to NGS techniques (Mróz, 2019).

I.6. Aim of the study

The aim of this study was to characterize the phenotype of each mutant (S2 and S3 lines) as well as the differential expression of the genes of the mutants compared to the wild type B10. The goal was to understand the mechanisms underlying the differences between the phenotypes. The RNA-seq results were already obtained previously and needed to be verified by qPCR. Only about twenty genes in each line were selected for such analysis. Protein networks were built to help to choose the most interesting genes following the central place of the corresponding proteins in the network, their function and their level of expression. Finally, the DEGs were mapped on the cucumber chromosomes and heatmaps were created to have a general view of the DEGs in each line. A promoter region analysis of the DEGs was made to understand how they are regulated.

II. MATERIALS AND METHODS

II.1. Plant material

II.1.1. *In vitro* culture

The plants used are cucumber somaclonal lines S2 and S3 that were regenerated *in vitro*. S2 line was regenerated from a leaf explant of the cultivar "Borszczagowski" (a highly inbred homogenous line named B10, the wild type), by leaf callus regeneration (LCR) on a solid MS (Musharige and Skoog, 1962) medium containing 2,4-D: 0,8 mg/L and IPA: 0,6 mg/L during 16 to 18 weeks. S3 line was regenerated by cytokinin-dependent embryogenic suspension (CES) from a shoot apical bud of the same cultivar that was first put on a solid MS medium with 2,4D for 3 weeks. Then the gel-like callus obtained was transferred in a cytokinin-dependent suspension (CDS) medium containing BAP: 2 mg/L during 9 to 10 months. Both lines were obtained in 1996. They were then self-pollinated around 10 times every 2-3 years and stored in a gene bank under controlled conditions. They are expected to be homozygous.

II.1.2. Greenhouse growth

The seeds of S2 and S3 somaclones obtained from *in vitro* culture and the wild type B10 line were sown in pots in 6 replicates and grown in the greenhouse of Wolica field in Warsaw for 5 weeks from March to April 2018 with the aim of studying the potential phenotypical differences between the mutants and the wild type. The photoperiod was 16 hours of day and 8 hours of night. The temperature during the day was between 25 and 27°C and during the night, it was between 18 and 20°C. The light intensity was maintained around 1500 $\mu mol.m^{-2}s^{-1}$.

II.2. RNA isolation and cDNA synthesis

RNA of the wild type B10 and of the 2 somaclonal lines S2 and S3 was extracted from plants that were cultivated in the field from June to July 2014, in 3 biological replicates each. The plants were irrigated. The mean temperature was around 18°C. 10 days after fertilization, the fruits were harvested and samples were ground and frozen in liquid nitrogen, then stored at -80°C. Total RNA was isolated from the samples with the miRNeasy Mini Kit (Qiagen, USA), following the manufacturer's instructions, including the optional steps (especially the DNase treatment). The concentration of the extracted RNA was measured, and its purity was verified by measuring the 260/280 and 260/230 ratios with a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The RNA was run on a 1% agarose gel, buffer TAE 1X, at 100 Volts for 30 minutes, with the intercalating agent ethidium bromide and revealed by UV light, to check that it was not contaminated by DNA. Once the concentration determined and the purity established, the RNA was diluted to a concentration of 100 ng/L. This RNA was also used previously to this work in the same project supported by a grant from the National Science Center (2013/11/B/NZ9/00814) for RNA-seq analysis.

cDNA was synthesized from the extracted RNA with the High-Capacity Reverse Transcription Kit (Thermo Fisher Scientific, Waltham, MA, USA) following manufacturer's instructions. The program of the thermocycler for cDNA synthesis is described in the **Annex I.1**. The cDNA obtained had a concentration of 100 µM. Then, it was diluted to get a concentration of 20 µM. The purity and the presence of cDNA were checked by PCR using the

primers of *UBIep*, a reference gene selected based on a previous work (Skarzyńska *et al.*, 2016). For the PCR reaction, 60 ng of cDNA were used (20 ng of each biological replicate) in a PCR mix of 21 µL containing 13,8 µL of MQ water, 2,5 µL of DreamTaq Buffer, 2,5 µL of dNTP (2mM), 1 µL of each Reverse and Forward primers (10 µM) and 0,2 µL of DreamTaq Polymerase for 1 sample. The PCR reaction is described in the **Annex I.2**. The products of the PCR were run on a 1% agarose gel, buffer TAE 1X, at 100 Volts for 30 minutes, with the intercalating agent ethidium bromide and revealed by UV light.

II.3. Bioinformatic network modeling of DEGs

The protein networks were created with STRING algorithm (version 11.0) (Szklarczyk *et al.*, 2015). The list of DEGs, previously established in the project supported by a grant from the National Science Center (2013/11/B/NZ9/00814) for each somaclonal line studied, was used to create the input files. The list of DEGs was established through an Illumia RNA-seq analysis where the genes were considered being differentially expressed compared to the wild type B10 when their fold change was greater than 1,5 and if the false discovery rate (FDR) was lower than 0,001. First, the sequences of the DEGs were put in Blast2Go (Götz *et al.*, 2008) to obtain the correct identifiers for cucumber's proteins. The same sequences were also put in Blast2Go (Götz *et al.*, 2008) to obtain the names of the corresponding proteins in *Arabidopsis thaliana*. The files were sorted out so for one query, the best result according to the percentage of identity, the bitscore and the e-value appears first. The input files for each somaclonal line were created by picking up the best result for each DEG. Several kinds of networks were created but only one was selected as it was more precise. The four different tested ways of building the network were : the DEGs' sequences as an input file with *Cucumis sativus* as a model, the same sequences with *Arabidopsis thaliana* as a model, the identifiers of cucumber's proteins as an input file and the identifiers of *Arabidopsis* proteins as an input file. The selected one was the identifiers of *Arabidopsis* as an input. The minimum confidence for the predicted network was set at 0,4. The layout of the network was edited with Cytoscape stringAPP (version 3.7.1) (Shannon et al., 2003). The color of each node (a node represents a protein) was attributed following the fold change of the corresponding gene obtained in the RNA-seq analysis. The downregulated genes were colored in green while the upregulated genes were colored in red. The darker the color is, the higher the fold change is. The thickness of every edge that connects the proteins shows the strength of the link between the proteins. The thicker the edge is, the stronger the connection is.

The networks were used to select the genes to verify by qPCR among the DEGs. The most important criterions were in this order: the place in the protein network, the function and the fold change. The selected genes are, in the network, the most central nodes, with the highest differential expression and which function is known and considered as important and interesting to study. For instance, these functions can be linked to hormones, lipids, cell wall, etc.

II.4. Verification of RNA-seq data by qPCR

The expression of the genes selected through the protein networks was verified by qPCR. The primers were designed with the Primer3Plus software (Untergasser *et al.*, 2007) and the sequence of *Cucumis sativus* B10 wild type line. The pair with the lowest interaction (with

itself and between both primers of the pair) was selected among the possibilities offered by the software, also taking into account the melting temperature that should be the most similar possible between the two primers of one pair. The self- and cross-interactions as well as the hairpins were verified and calculated by the Beacon Designer™ Free Edition software. The less interactions and hairpins the pair of primers has, the better it is. This software also gives the number of GC clamp which is important for the primers' specificity. Selected primers usually have one or two GC clamps. Therefore, the specificity was checked with Primer-BLAST, the primer designing tool of NCBI (https://www.ncbi.nlm.nih.gov/tools/primer-blast/). It verifies the specificity of the area of the gene where the primers bind, but also the specificity of one isoform compared to the others by giving all the possible binding areas on the gene and all its isoforms. Thus, the primers were chosen to be as specific as possible.

The primers were diluted according to the manufacturer's instruction to obtain a concentration of 100µM. Working dilutions were made to prepare the qPCR reactions. The concentration needed is 10µM. The primers were tested by PCR with the same mix and same cDNA as it was described before and the same program that is described in the **Annex I.2**. The products were run on a 1% agarose gel, buffer TAE 1X, at 130 Volts for 30 minutes to 1 hour, with the intercalating agent ethidium bromide and revealed by UV light. When no band or more than one band were detected, the primers were tested on a gradient PCR from 50 to 65°C. The mix and cDNA were the same as previously described. The PCR program used is described in **Annex I.3**. The products of the PCR were run on a 1% agarose gel, buffer TAE 1X, at 130 Volts from 30 minutes to 1 hour, with the intercalating agent ethidium bromide and revealed by UV light. One band was expected. When there were more bands or no bands, the primers were not used in the qPCR reaction.

The expression of the genes that showed correct products on the PCR or the gradient PCR was verified by qPCR. From all the DEGs in both somaclonal lines, 16 of S2 line and 17 of S3 line were analyzed by qPCR. The list of these genes is available in **Annex II**. Moreover, 2 reference genes, *UBIep* and *TIP41* were selected from a previous work (Skarzyńska *et al.*, 2016) through a PCR with the same mix and same cDNA (only from S2 and B10 lines) as it was described before and the same program which is described in the **Annex I.2**. It was assumed that these reference genes were suitable for the analysis with S3 line as well. Three biological replicates with 2 technical replicates were used. For the qPCR reaction, 20 ng of cDNA was used in each sample of a qPCR mix of 11 µL containing 2,3 µL of MQ water, 7,5 µL of Power SYBR® Green PCR Master Mix (Thermo Fisher Scientific, Waltham, MA, USA) and 1 µL of each Reverse and Forward primers (10 µM) for 1 sample. Three biological replicates with 3 technical replicates each were used. The qPCR reaction is described in the **Annex I.4**. Directly after the qPCR reaction, a melting curve analysis was done.

II.5. qPCR data treatment

The raw data obtained by the qPCR reaction was formatted in RStudio through a script to be entered, one gene after another, in the software LinRegPCR (version 2017.1) (Ramakers *et al.*, 2003). It determined the baseline and calculated the mean qPCR efficiency for each gene basing on the linear regression of the slope's regression line in the exponential stage. Then, the relative expression of each DEG in the somaclonal lines was calculated following a modified and more accurate $2^{-\Delta\Delta Ct}$ method on Rstudio through a script with EasyqPCR from the

Bioconductor software package (Hellemans *et al.*, 2007). They were normalized to the wild type B10 and to the reference genes, *UBIep* and *TIP41*. The standard deviations were calculated on RStudio with the same script. The charts were elaborated following the results of the data analysis on Excel.

II.6. Heatmap construction

The transcripts per million (TPM) of the genes of each replicate of each somaclonal line and the wild type B10 were used as an input file in the software MeV 4.9.0 (http://mev-tm4.sourceforge.net/svnroot/mev-tm4/trunk) to build a heatmap. The log10 of the TPM was taken and then the data were normalized. Finally, the genes were clustered using hierarchical clustering (HCL) method with Pearson correlation statistics.

II.7. Bioinformatics analysis of the patterns in the gene's promoters

The list of the DEGs promoters' sequences was used as an input file on PlantCare (Rombauts *et al.*, 1999). From the results given by the software, an Excel file was created giving the functions of the promoter region for each gene in several organisms and how many times the function appears in this combination of gene and organism. The charts were elaborated on Excel.

II.8. DEGs' location on chromosomes

A chromosome map of the DEGs was built using the MapChart software (Voorrips, 2002). A database containing the list of contigs in *Cucumis sativus*, the chromosome where they are located and other information on the contigs was used. However, some contigs were mapped on several chromosomes, and some contigs are not yet mapped. Therefore, the latter ones were omitted for the mapping as well as the DEGs that were mapped on them. Moreover, several hypotheses were made. To determine on which chromosome it is more likely to find each contig, it was assumed that the probability of finding a contig in a specific chromosome was all the more high as the quotient of the number of markers on this chromosome for this contig by the total number of markers for this contig is high. That is to say, the higher the proportion of markers for the contig on one chromosome is, the more likely it is that the contig is located on this chromosome. Then, to locate more precisely the contig on the chromosome, a second hypothesis was made, saying that the location of the contig on the chromosome is equal to the average position of the markers for this contig. It was also supposed that the length of the chromosome is equal to the sum of the length of all the contigs it contains, despite that it is possible that the contigs are overlapping or that they have spaces between them, and knowing that 15 contigs for S2 and 46 contigs for S3 are missing because they are not mapped yet.

III. RESULTS

III.1. Phenotypes of the somaclonal lines

The vegetative parts of *Cucumis sativus* plants of the line S2 are quite similar to the ones of the wild type B10 (**Fig. 2a, b, c**). They have the same shape and color of leaves. The stage of growth is the only thing that can differentiate them. Indeed, a delay of growth can be noticed (**Fig. 2b**). In the first four weeks after sowing, the plants of S2 line remain smaller than the ones of B10 line. However, on the fifth week, they reached the same height, but then a delay of development was noticeable through a delay in flowering. B10 plants already had few open flowers while S2 plants' flowers were still in formation (**Fig. 2c**).

Cucumis sativus S3 line has a noteworthy delay of growth, remaining half-size of B10 line plants at every stage (**Fig. 2d, e**). It might be a dwarf form of cucumber plant. Yet the leaves reach the normal size around the fourth week of culture after sowing. Moreover, the growth habit of these plants is more compact. The biggest difference between the wild type and the S3 mutants lies in the color of the leaves. Young leaves are light green-yellowish with very marked veins of the same color and have a slightly different shape (**Fig. 2d, e**). They are more elongated. As for the old leaves, they become darker, the same color as B10 leaves. However, their veins remain light green-yellowish, making S3 plants easy to distinguish from B10 plants.

III.2. Statistics on differentially expressed genes

The DEGs identified were considered to be differentially expressed if their fold change was higher than 1,5 with a false discovery rate lower than 0,001. In S2 line, 364 DEGs were identified in which 324 are protein coding (89% of the DEGs) and 40 are long intergenic non-coding RNA (lincRNA) (11% of the DEGs). Ninety DEGs were downregulated (25%) and 274 were upregulated (75%). In S3 line, 273 DEGs were identified in which 203 are protein coding (74% of the DEGs) and 67 are lincRNA (25% of the DEGs). The 1% left goes to small nucleolar RNA (snoRNA) of CD-box type, that is responsible for the methylation of ribose of the rRNA, participating in the rRNA maturation. S2 and S3 have 26 common DEGs (**Fig. 3**). One hundred nine DEGs were downregulated (40%) and 164 were upregulated (60%)

III.3. Verification of RNA-seq results by qPCR

To validate the RNA-seq results, qPCR reactions were carried out using specific primers for the chosen DEGs of S2 and S3 lines. In order to view the qPCR results and RNA-seq data simultaneously, a set of charts was built (**Fig. 4a, b**). Each chart represents one gene. Seeing the trend of the difference of expression between B10 and S2, and between B10 and S3, the RNA-seq data are validated for all the genes in both somaclonal lines except one (*Cucsat.PASA.G149*) that has overlapping error bars (**Fig. 4b**). Despite the correlated trend of the qPCR results of this gene with the RNA-seq data, this result cannot be validated because it is not significant. However, 100% of S2 line and 94% of S3 line DEG expression profiles is validated by qPCR.

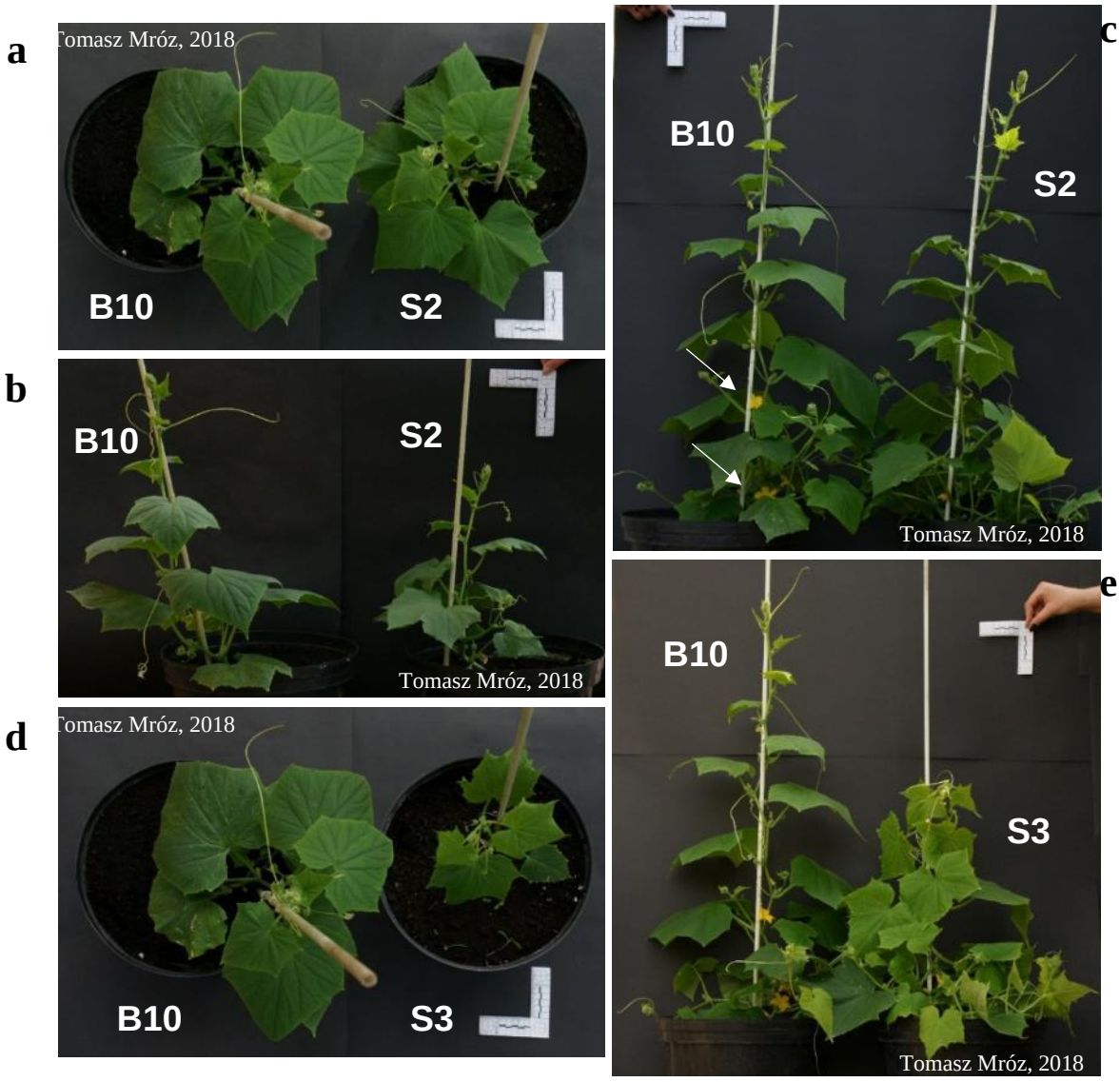


Figure 2: Photographs showing the comparison between the somaclonal lines S2 and S3 and the wild type B10 from different angles and at different stages of growth. S2 is compared to B10 after **a)** 4 weeks of growth, from the top **b)** from the side and **c)** after 5 weeks of growth, from the side. S3 is compared to B10 **d)** after 4 weeks of growth, from the top **e)** after 5 weeks of growth, from the side. The white arrows show the flowers

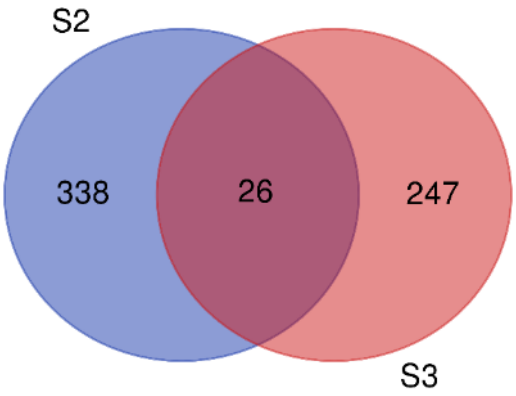


Figure 3: Venn diagram presenting the number of DEGs in S2 and S3 line and the number of common DEGs between the two somaclonal lines

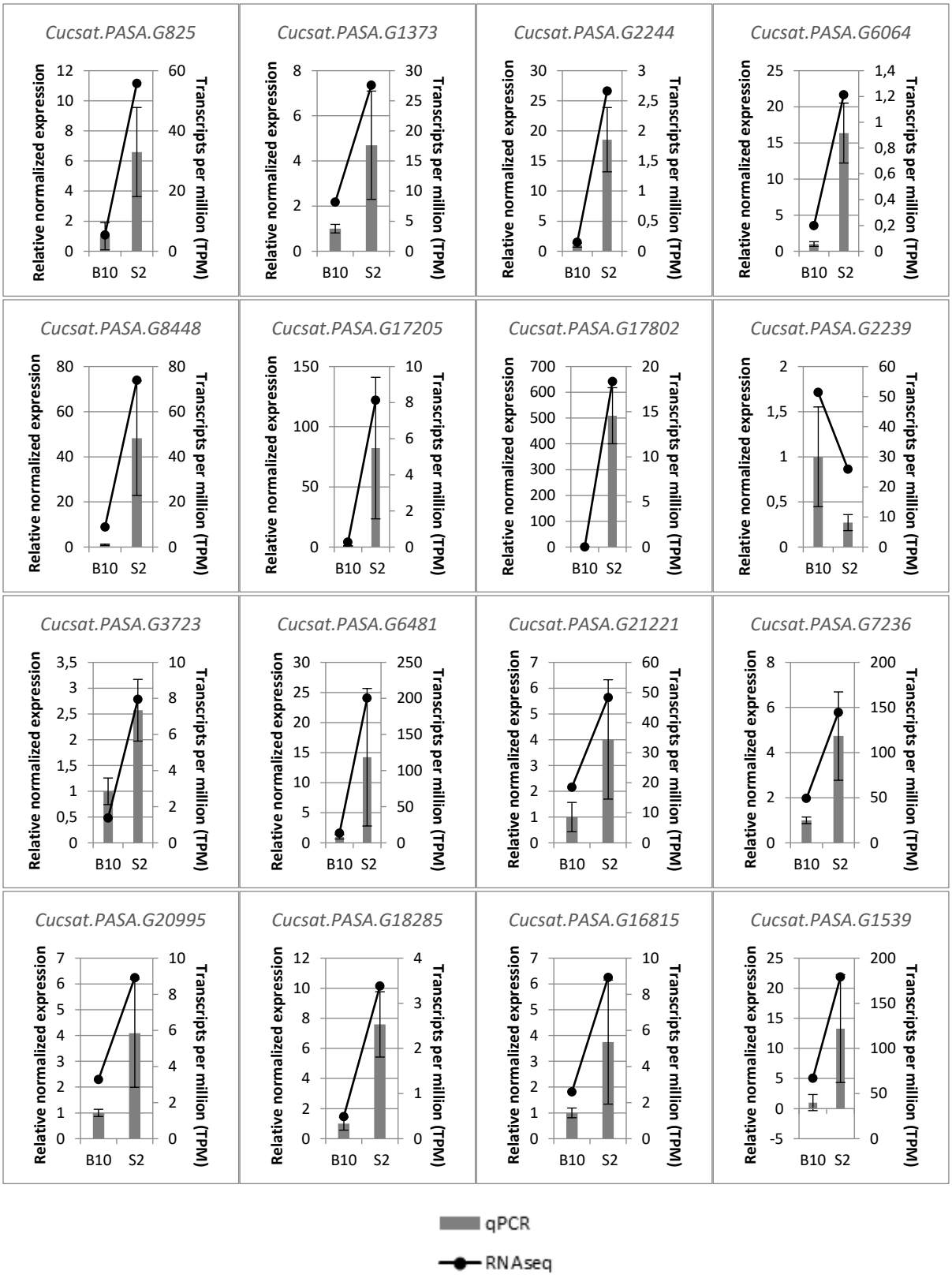


Figure 4a: RNA-seq data, expressed in TPM and represented by the black line, in comparison with the qPCR results, expressed as the relative normalized expression of the S2 DEGs and represented by the grey bars. Each chart corresponds to 1 of the 16 selected DEGs of S2 line. The error bars represent the standard deviation of the three biological replicates

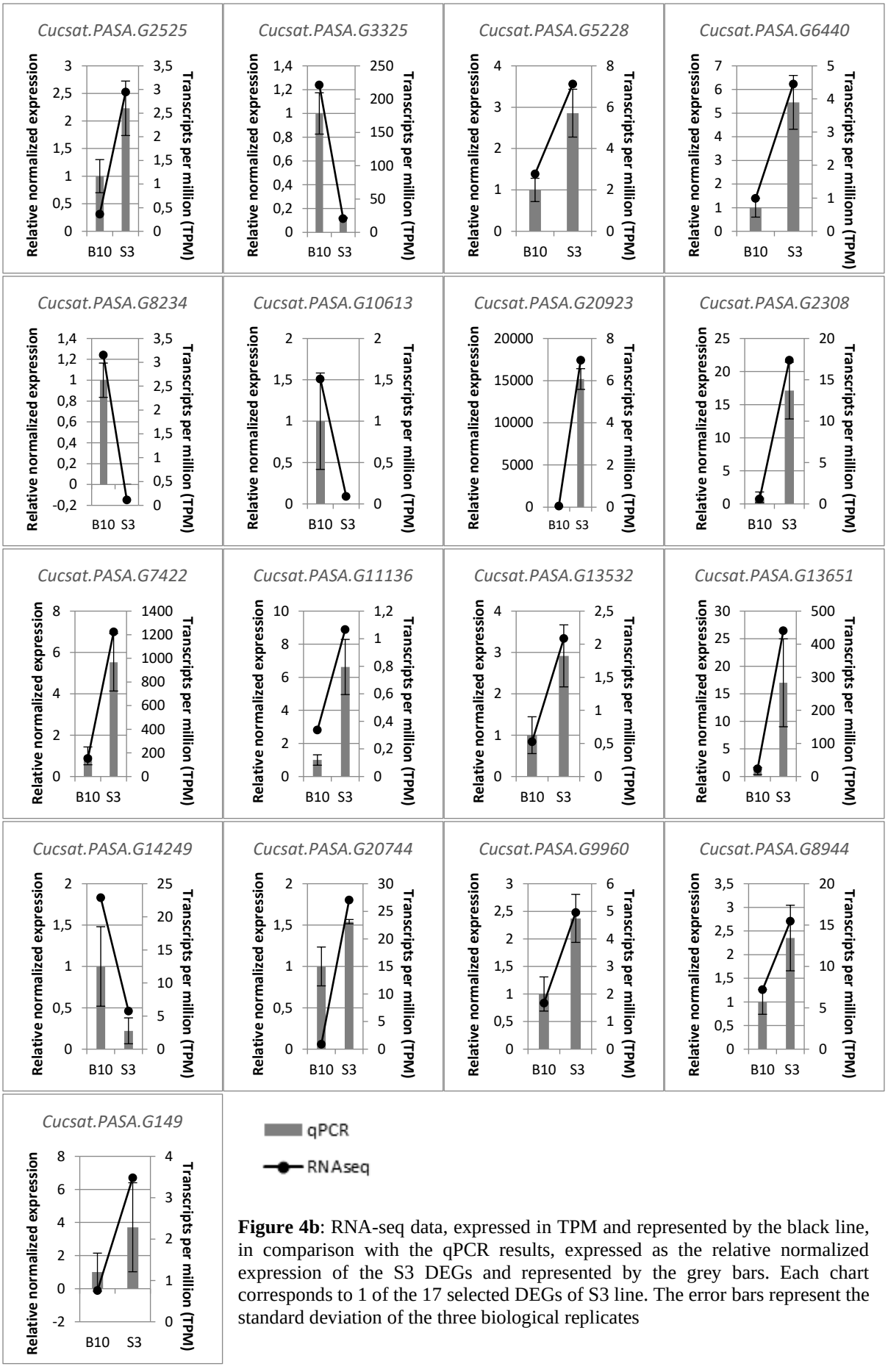


Figure 4b: RNA-seq data, expressed in TPM and represented by the black line, in comparison with the qPCR results, expressed as the relative normalized expression of the S3 DEGs and represented by the grey bars. Each chart corresponds to 1 of the 17 selected DEGs of S3 line. The error bars represent the standard deviation of the three biological replicates

III.4. Analysis of protein interaction through bioinformatics tools

Two molecular networks were built based on STRING analysis of the DEGs of S2 and S3 somaclonal lines.

The S2 network is made of 306 nodes and 200 edges (**Fig. 5a**), with an average node degree of 1,31. The average node degree is the average number of interactions that a protein has in the network. The average local clustering coefficient is of 0,279. This coefficient represents a measure showing how connected are the nodes of the network. The higher the value is the stronger the nodes are connected to each other. Moreover, the protein-protein interaction (PPI) enrichment p-value is lower than 1.0e-16, indicating that the number of edges, the enrichment of the network, is significant, and the proteins are not random but belong to groups. Besides, 144 nodes were identified as single components that are not implied in any network, and 6 nodes were identified as forming pairs. Five other networks containing each 3 nodes and 2 edges, one network containing 3 nodes with mutual interactions, one network containing 4 nodes and 4 edges, and one network containing 5 nodes and 5 edges were identified. The main network is made of 129 nodes and 175 edges. In this main network, 4 groups were identified, explaining the significant enrichment value: 1. proteins implied in transcription, 2. proteins related to cell wall and others, 3. protein related to lipids, 4. intricate group of proteins of various functions related to hormones, detoxification, amino acids, transporters, transcription, proteolysis, vesicle trafficking, metal ions, stress response and others.

The S3 network is made of 186 nodes and 65 edges (**Fig. 5b**), with an average node degree of 0,699. The average local clustering coefficient is 0,281. Moreover, the PPI richment p-value is 0,00139. The p-value is higher than the one of S2 network but is still low enough to consider that the network's enrichment is significant and that the proteins are not random but belong to groups. Besides, 108 nodes were identified as single components that are not implied in any network, and 24 nodes were identified as forming pairs. Two other networks containing 3 nodes and 2 edges, one network containing 4 nodes and 3 edges, 1 network containing 5 nodes and 4 edges and one network containing 9 nodes and 9 edges were identified. The main network is made of 30 nodes and 33 edges. In this main network and the smaller network of 9 nodes, 7 groups were identified, explaining the enrichment value (**Fig. 5b**): 1. proteins implied in cell cycle processes, 2. proteins implied in transcription, 3. proteins implied in translation and maturation of proteins, 4. proteins implied in flowering processes, 5.a. ions and b. other transporters, 6. proteins implied in the peroxisome metabolism, 7. proteins implied in detoxification processes.

III.5. Bioinformatics analysis of promoter regions of DEGs in somaclonal lines

After determining the DEGs, their expression, their functions and interactions, it is interesting to understand how they are regulated, by which kind of factors. To reach this aim, the PlantCare database was used to study and analyze the promoters of the DEGs. The results were organized in charts representing the number of elements in the different categories of motifs found, the different lengths of the motifs' sequences, the number of elements in each organism containing these motifs. The functions of the motifs were manually grouped together and the number of elements possessing each category of function is also represented on a chart.

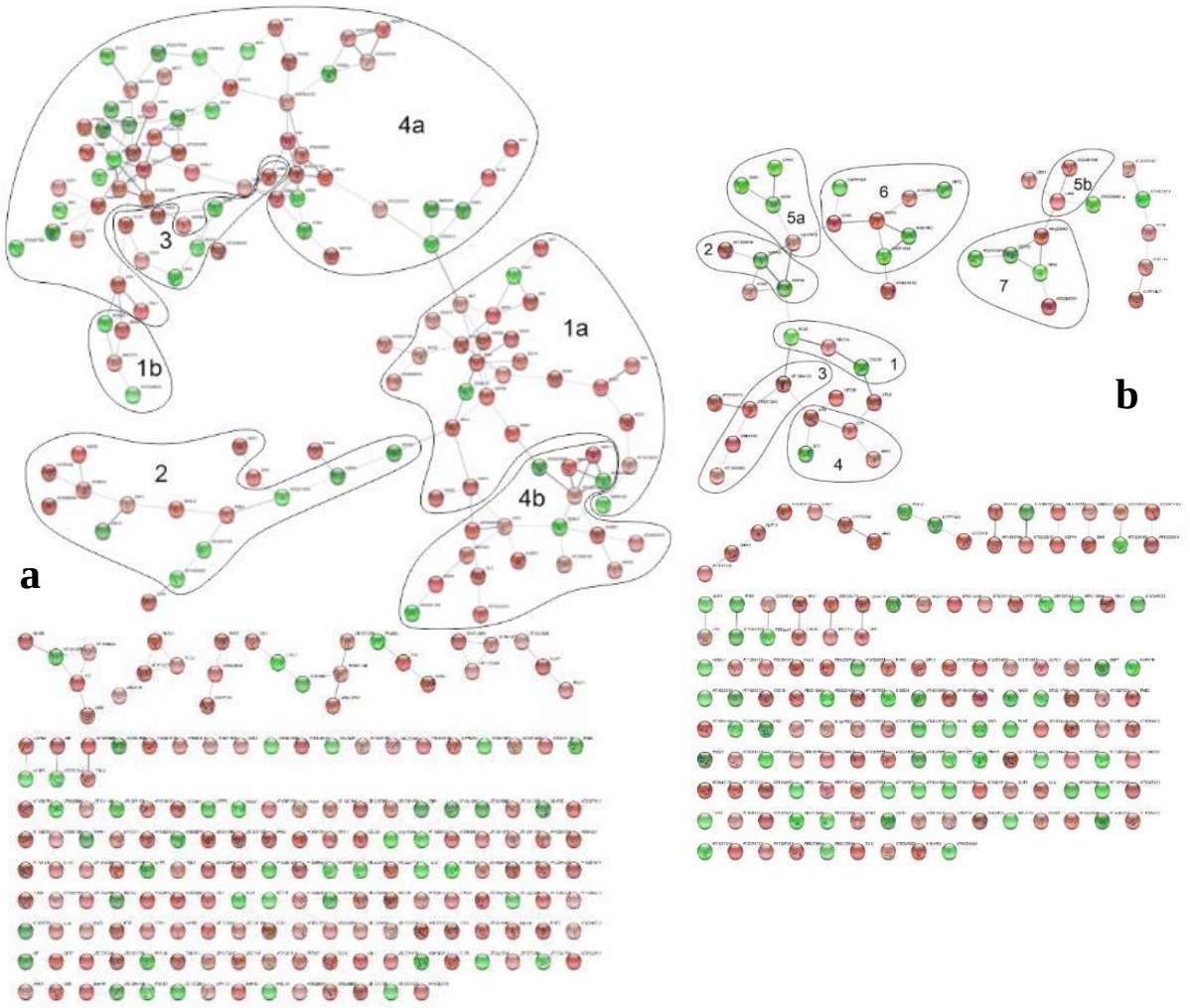


Figure 5: Protein network showing the connections between the DEGs of **a)** S2 line and **b)** S3 line. Each node represents one protein and each edge a connection. The red color means that the gene is upregulated and the green color means that the gene is downregulated. The darker the color is, the higher the fold change is. The thicker the edge is, the stronger the connection is. In S2 line, the main network was divided in the 4 following groups: 1. proteins implied in transcription, 2. proteins related to cell wall and others, 3. protein related to lipids, 4. intricate group of proteins of various functions related to hormones, detoxification, amino acids, transporters, transcription, proteolysis, vesicle trafficking, metal ions, stress response and others . In S3 line, the two main networks were divided into the 7 following groups: 1. proteins implied in cell cycle processes, 2. proteins implied in transcription, 3. proteins implied in translation and maturation of proteins, 4. proteins implied in flowering processes, 5.a. ions and b. other transporters, 6. proteins implied in the peroxisome metabolism, 7. proteins implied in detoxification processes

In the PlantCare analysis of the promoter region of the S2 DEGs, 24158 motifs in different organisms, with different sequences and functions were found. They are divided in 116 motifs. In the **fig. 6a**, only the 47 most abundant of them are showed. These all appear at least 30 times in the analysis. The motifs appearing less than 30 times are grouped under the name “Others”. The most represented motifs are the TATA-box and CAAT-box, appearing respectively 7976 and 6008 times and constituting respectively 33% and 25% of the total number of analyzed sequences. There are still a lot of unknown motifs of which sequence is not known yet. They appear 1441 times and represent 6% of the analyzed sequences. Moreover, some motifs can be found under name types like “Unnamed_1”, “Unnamed_2”, etc. These motifs possess a known

sequence but are not named yet. One of these categories, "Unnamed_4" is quite highly represented in this analysis since it appears 1117 times and represents 5% of the sequences. The other motifs, going from the AT~TATA-box to ABRE, represent each between 1% to 5% of the analyzed sequences while each of all the others represent less than 1% of the analyzed sequences.

In the PlantCare analysis of the promoter region of the S3 DEGs, 17528 motifs in different organisms, with different sequences and functions were found. They are divided in 109 motifs. In the **fig. 6b**, only the 45 most abundant of them are showed. These all appear at least 30 times in the analysis. The motifs appearing less than 30 times are grouped under the name "Others". The most represented motifs are the TATA-box and CAAT-box, appearing respectively 5472 and 4256 times and constituting respectively 31% and 24% of the total number of analyzed sequences. In this line also there are still a lot of unknown motifs. They appear 1057 times and represent 6% of the analyzed sequences. The same category as in S2 line, the category "Unnamed_4", is quite highly represented in this analysis since it appears 893 times and represents 5% of the sequences. The other motifs, going from the AT~TATA-box to CGTCA-motif, represent each between 1% and 5% of the analyzed sequences while each of all the others represent less than 1% of the analyzed sequences.

The length of the motifs in S2 line is between 4 and 18 nucleotides (**Fig. 7a**). The length of some motifs was measured equal to a non-entire number (8,5; 9,5; etc.). It means that the sequence can sometimes count one more nucleotide, but the nature of this nucleotide is unsure. The most recurrent motifs are short, between 4 and 6 nucleotides. They are 20340 which makes 84% of the total analyzed sequences. The motifs over 10 nucleotides and the ones counting a non-entire number of nucleotides are weakly represented. They are 187 and represent together less than 1% of the analyzed sequences, as well as the motifs made of 10 nucleotides (199). As for the motifs between 7 and 9 nucleotides, each of them represents between 4% and 6% of the analyzed sequences.

The length of the motifs in S3 line is between 4 and 15 nucleotides (**Fig. 7b**). The most recurrent motifs are short, between 4 and 6 nucleotides. They are 14916 which makes 85% of the total analyzed sequences. The motifs over 10 nucleotides and the one counting a non-entire number of nucleotides are weakly represented. They are 154 and represent together less than 1% of the analyzed sequences, as well as the motifs made of 10 nucleotides (135). As for the motifs between 7 and 9 nucleotides, each of them represents 4% or 5% of the analyzed sequences.

The motifs were found in 31 different organisms for S2 line while some motifs are still not attributed to any organism, constituting the "Unknown" category (**Fig. 8a**). The motifs were mainly attributed to the model plant *Arabidopsis thaliana*. It represents 9424 of the motifs, that being 39% of the analyzed sequences. *Nicotiana glutinosa, Pisum sativus* and *Brassica napus* were also quite well represented with respectively 3489, 2440 and 1244 of the motifs, which makes respectively 14%, 10% and 5% of all the analyzed sequences. There are 1441 motifs that are not attributed to any organism, that is to say 6% of the analyzed sequences. The motifs of S3 line were found in 30 different organisms (**Fig. 8b**). There is an "Unknown" category in this line as well. The motifs were mainly attributed to the model plant *Arabidopsis thaliana*. It represents 6626 of the motifs, that being 38% of the analyzed sequences. *Nicotiana glutinosa*, *Pisum sativus*, *Brassica napus* and *Petroselinum Hortense* were also quite well

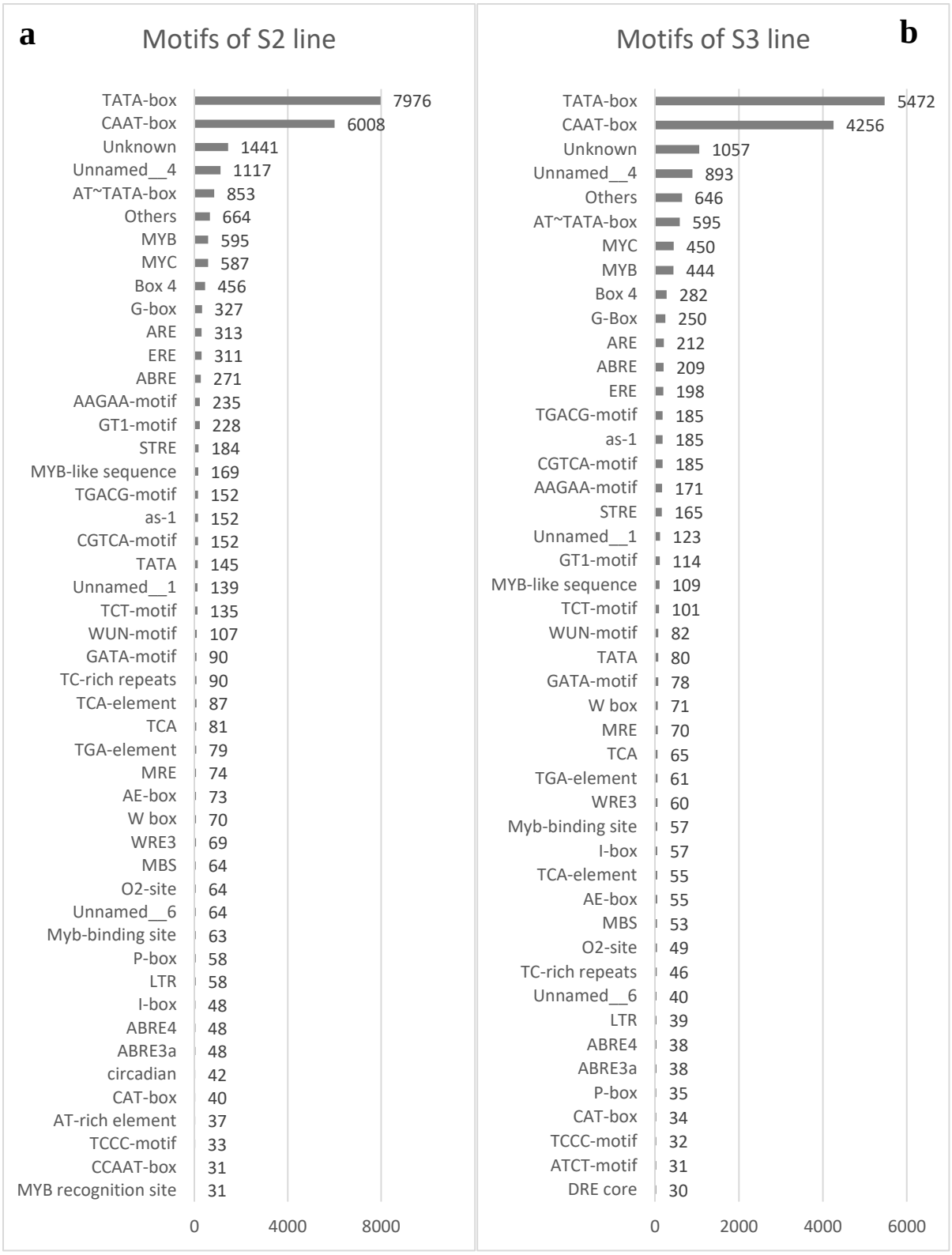


Figure 6: Abundance of each motif found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line. Only the motifs present more than 30 times in the total analyzed sequence are shown

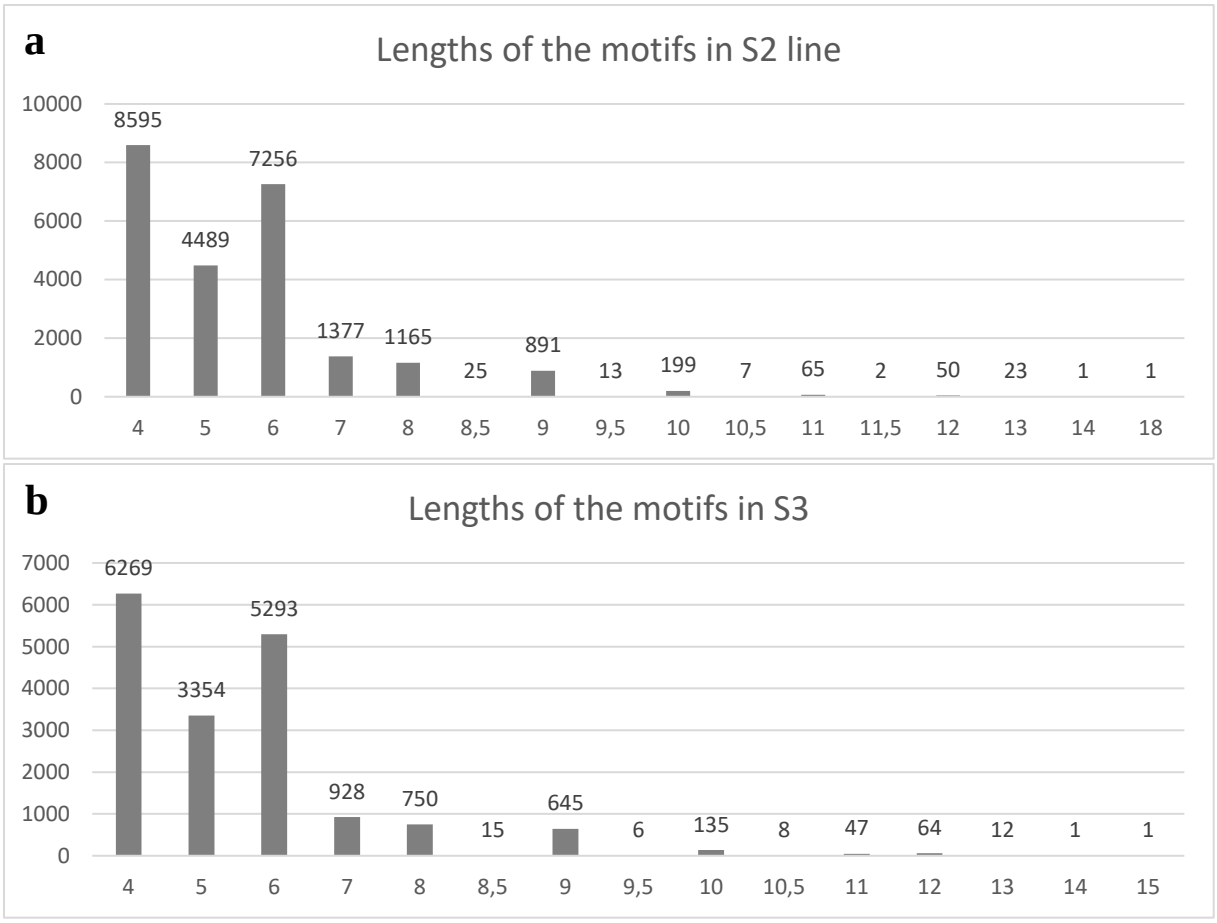


Figure 7: Distribution of the motifs' length found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line. The non-entire lengths mean that the motif could have one more nucleotide which nature is unsure

represented with respectively 2585, 1701, 945 and 937 of the motifs, which makes respectively 15%, 10%, 5% and 5% of all the analyzed sequences. There are 1057 motifs that are not attributed to any organism, that is to say 6% of the analyzed sequences.

For S2 line, the functions were manually classified in 18 categories of function (**Fig. 9a**). Most of the functions remain unknown. There are 12806 motifs in this category, constituting 53% of the analyzed sequences. Otherwise, the most represented function is the core promoter elements with 7976 motifs in this category, that being 33% of the analyzed sequences. The light response elements is the third most represented function with 1693 motifs and 7% of the analyzed sequences. Hormone, stress and oxygen response elements are following, representing between 1% and 2% of the analyzed sequences. The last ones are, in the following order, the metabolism regulation elements, the circadian cycle regulation elements, the meristem expression elements, the endosperm expression elements, the cell differentiation elements, the flavonoid biosynthetic genes regulation, the seed-specific regulation elements, the cell cycle regulation elements, the phytochrome expression elements, the hormone and light response elements and the flowering regulation elements, each of them representing less than 1% of the analyzed sequences.

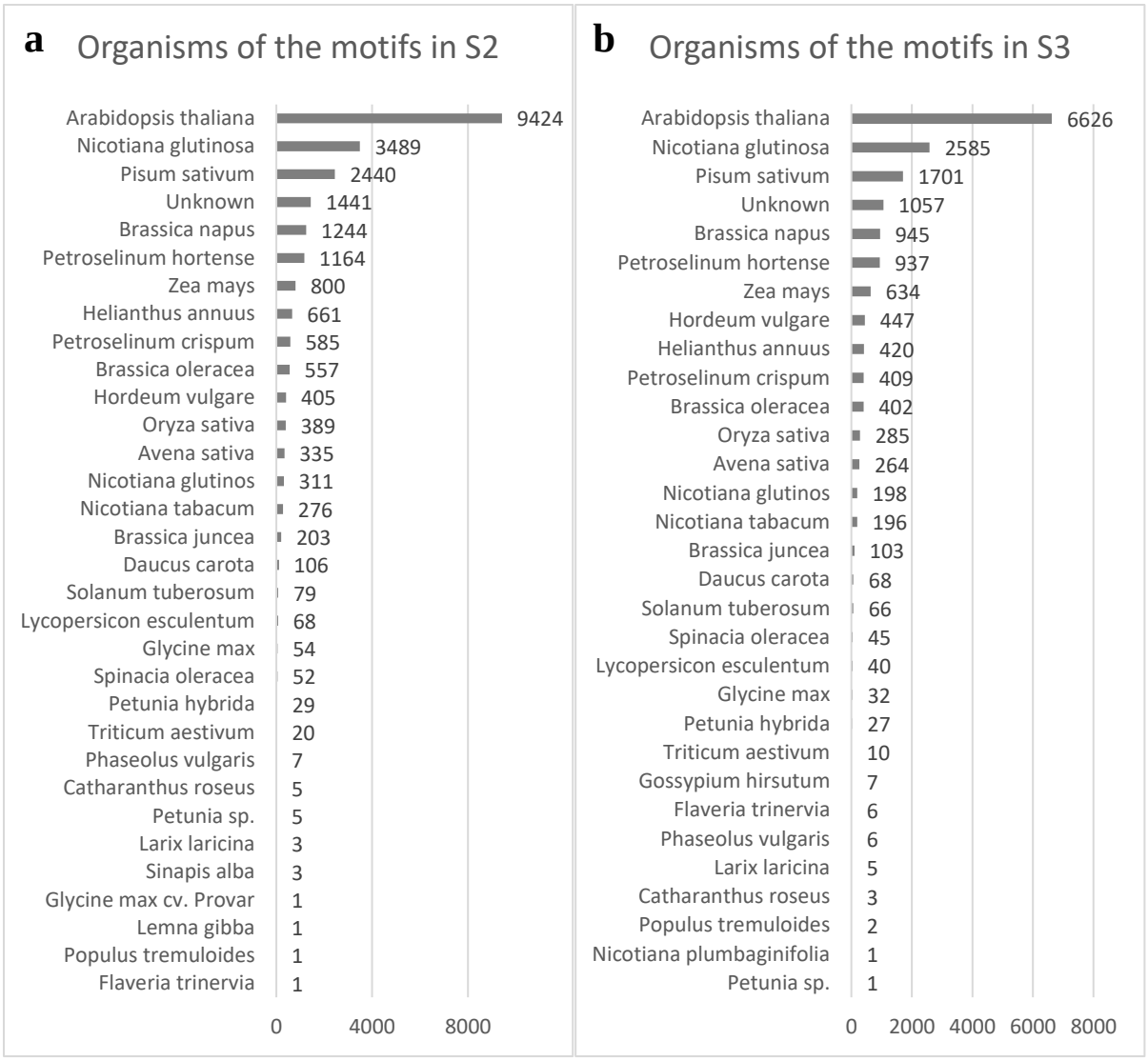
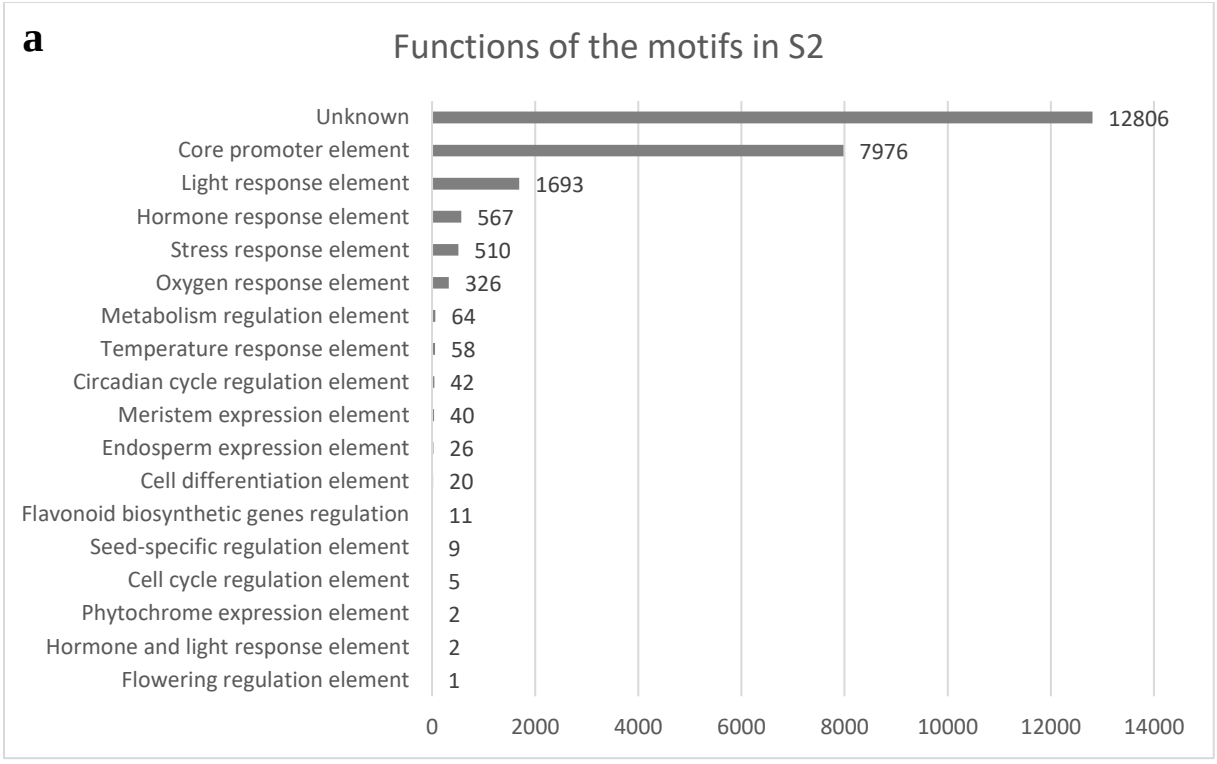


Figure 8: Distribution of the organisms to which the motifs found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line were attributed

For S3 line, the functions were manually classified in 17 categories of function (**Fig. 9b**). Most of the functions remain unknown. There are 9464 motifs in this category, constituting 54% of the analyzed sequences. Otherwise, the most represented function is the core promoter elements with 5472 motifs in this category, that being 31% of the analyzed sequences. The light response elements is the third most represented function with 1229 motifs and 7% of the analyzed sequences. Stress, hormone and oxygen response elements are following, representing between 1% and 3% of the analyzed sequences. The last ones are, in the following order, the metabolism regulation elements, the temperature response elements, the meristem expression elements, the circadian cycle regulation elements, the endosperm expression elements, the cell differentiation elements, the flavonoid biosynthetic genes regulation, the organ-specific regulation elements, the phytochrome expression elements, the cell cycle regulation elements and the hormone and light response elements, each of them representing less than 1% of the analyzed sequences.



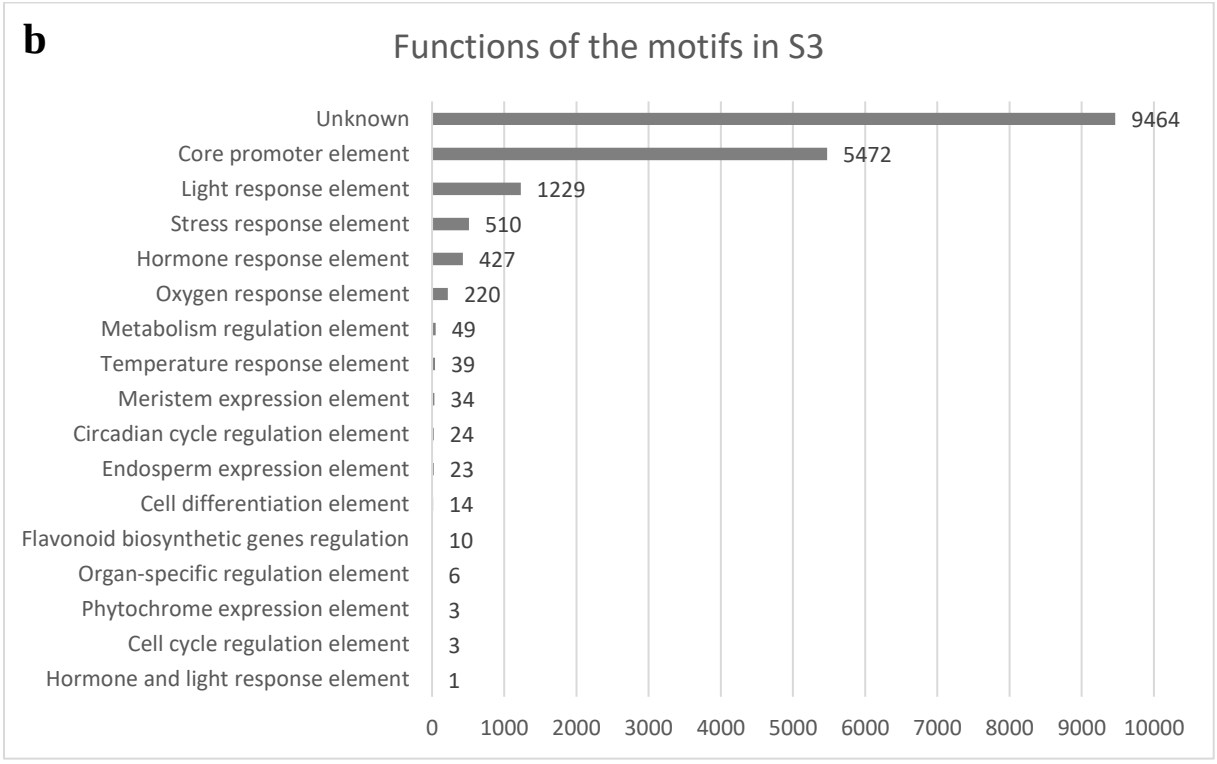


Figure 9: Distribution of the functions of the motifs found with PlantCare analysis in the promoter region of the DEGs in **a)** S2 line and **b)** S3 line. PlantCare functions were manually grouped in the presented functions

III.6. Chromosomal location of DEGs

The next step of the analysis consisted in determining the place of each DEG on the chromosomes of the cucumber. The places of the DEGs in their respective contigs and chromosomes are represented on chromosome maps on the **fig.10a** for S2 and on the **fig. 10b** for S3. The upregulated DEGs are represented in red while the downregulated DEGs are represented in green. The protein coding DEGs are written without any alteration of the font. The lincRNA are written in italic bold font and the snoRNA are underlined. In S2 line, 348 DEGs in total were mapped on the 7 chromosomes out of the 364 total DEGs. Some contigs are empty, some contain only 1 DEG while others contain whole groups of DEGs. Besides, the 3rd chromosome is the longest, but it is the 6th one that has the highest number of DEGs. Similarly, the 2nd chromosome is the 2nd shortest chromosome but has the 3rd highest number of DEGs while the 5th chromosome is the 2nd longest chromosome and has the 2nd lowest number of DEGs. Moreover, there is a great majority of protein coding genes among the mapped DEGs. Indeed, they are 314, that being 90% of these DEGs, while lincRNA are only 34 or 10% of these DEGs, and there is no snoRNA.

In S3 line, 223 DEGs in total were mapped on the 7 chromosomes out of the 273 total DEGs. Some contigs are empty, some contain only 1 DEG while others contain whole groups of DEGs. Furthermore, the 3rd chromosome is the longest, but the 6th one has the highest number of DEGs. Similarly, the length of all the other chromosomes except the 7th one does not correspond to the number of DEGs that are mapped on it. Moreover, there is a great majority of protein coding genes among the mapped DEGs. Indeed, they are 181, that being 81% of these DEGs, while lincRNA are only 39 or 18% of these DEGs and the snoRNA are 3, that is to say 1% of these DEGs. Besides, all the snoRNA are mapped on the 3rd chromosome.

III.7. Heatmap of DEGs

To complete the analysis of the DEGs, the repeatability of the biological replicates of the somaclonal S2 and S3 lines and the wild type reference line B10 was verified by building heatmaps (**Fig. 11**) that also enabled to better see the distribution of the differential expression of the DEGs. Genes were clustered following their differential expression by HCL analysis using the Pearson correlation coefficient and the gene tree built in this way was added on the left of the heatmap (**Fig. 11**). On S2 heatmap, most of the DEGs are expressed following the same trend within the three biological replicates. Thus, it can be considered that the differences between the biological replicates are not significant. Moreover, according to the HCL analysis, 2 main different anticorrelated groups can be distinguished. The anticorrelation is medium with a Pearson correlation coefficient of -0,27. Inside these 2 main groups many other moderately correlated groups with a Pearson correlation coefficient of 0,36 can be found (**Fig. 11a**).

On S3 heatmap, most of the DEGs are expressed following the same trend within the three biological replicates. Thus, it can be considered that the differences between the biological replicates are not significant. Moreover, according to the HCL analysis, 2 main different anticorrelated groups can be distinguished. The anticorrelation is medium with a Pearson correlation coefficient of -0,37. Inside these 2 main groups many other moderately correlated groups with a Pearson correlation coefficient of 0,32 can be found (**Fig. 11b**).

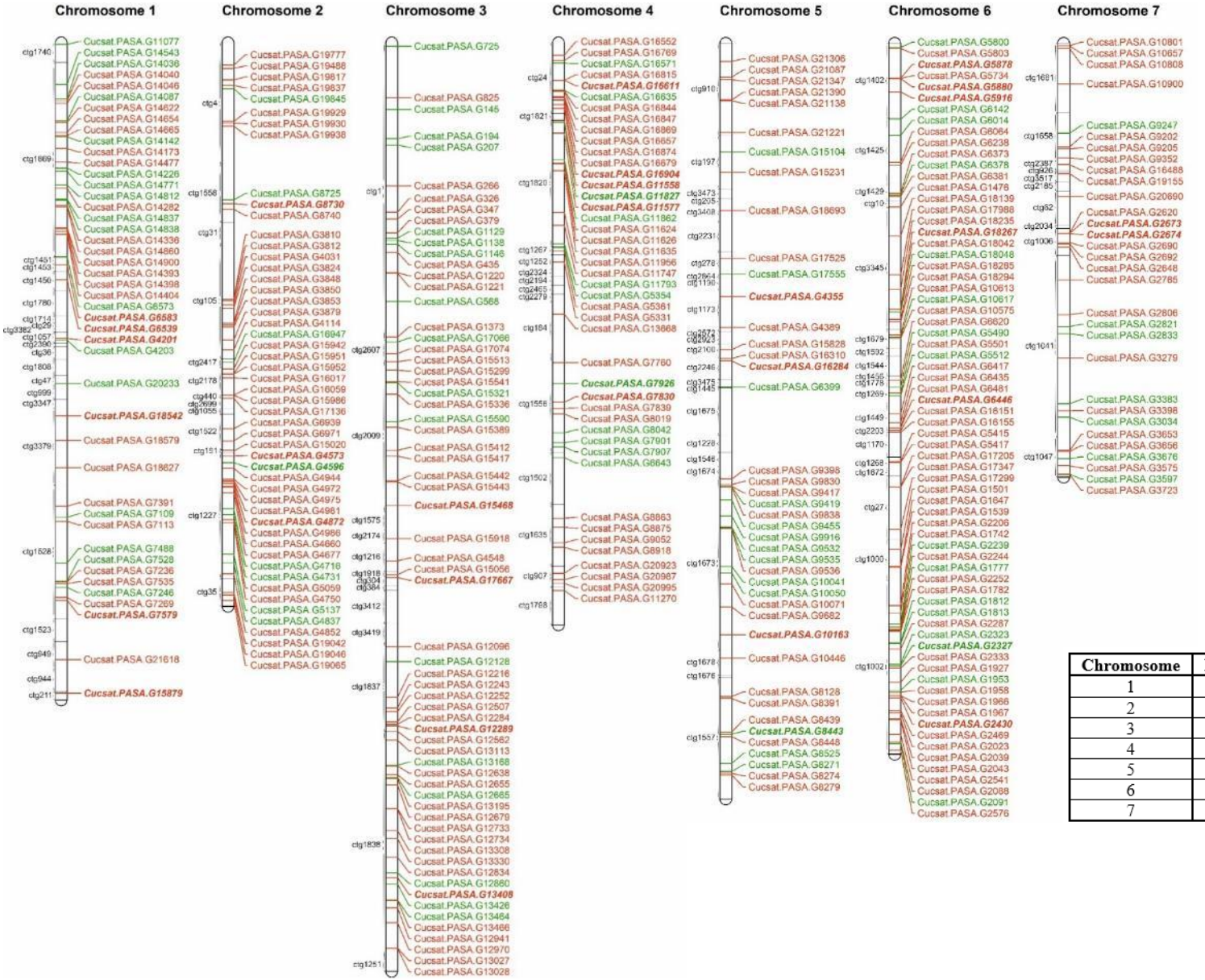


Figure 10a: Chromosome maps of all the cucumber chromosomes indicating the position of the DEGs of S2 line on the chromosome and more precisely on the contig. The contigs are delimited by black lines. The position of the gene is represented by a colored line. The red color means that the gene is upregulated and the green color means that the gene is downregulated. The names without any modification of the font correspond to the protein coding genes. The names that are in bold and italic font correspond to the lincRNA. The table gives the number of contigs and DEGs mapped on every chromosome

Chromosome	Number of contigs	Number of DEGs
1	22	45
2	13	51
3	14	65
4	15	44
5	24	42
6	19	70
7	11	31

Figure 10b: Chromosome maps of all the cucumber chromosomes indicating the position of the DEGs of S3 line on the chromosome and more precisely on the contig. The contigs are delimited by black lines. The position of the gene is represented by a colored line. The red color means that the gene is upregulated and the green color means that the gene is downregulated. The names without any modification of the font correspond to the protein coding genes. The names that are in bold and italic font correspond to the lincRNA. The underlined names correspond to snoRNA. The table gives the number of contigs and DEGs mapped on every chromosome

Chromosome	Number of contigs	Number of DEGs
1	22	21
2	13	31
3	14	51
4	15	18
5	24	34
6	19	52
7	11	16

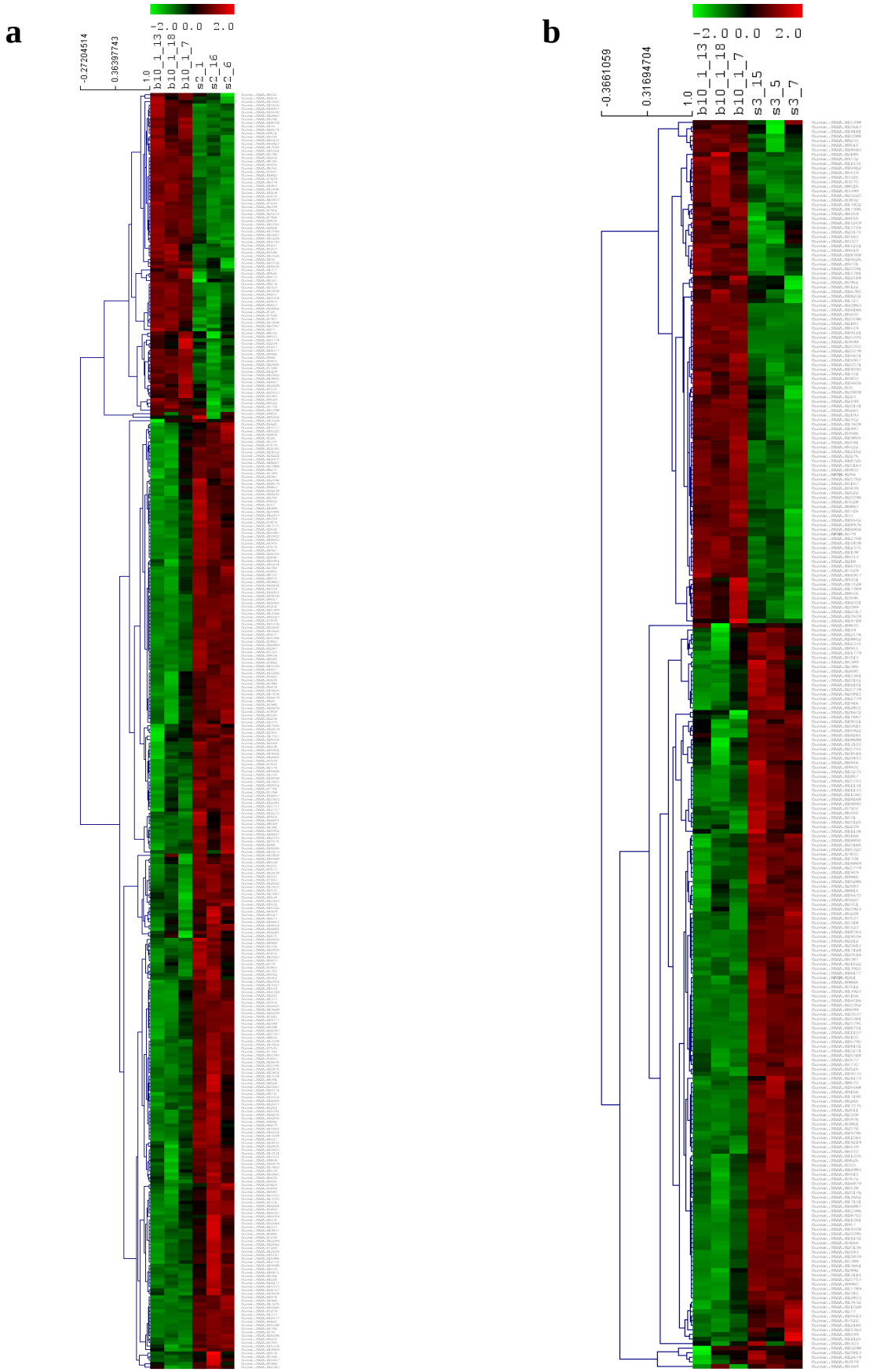


Figure 11: Heatmaps of **a)** S2 and **b)** S3 DEGs. Each heatmap shows the three biological replicates of the B10 wild type line and the three biological replicates of a somaclonal line. The HCL analysis results are shown on the left of each heatmap

IV. DISCUSSION

IV.1. Phenotypical analysis of the somaclonal lines of cucumber

The S2 line was defined to be identical to the wild type B10 except for its slower rhythm of growth implying a late flowering. The experiment was stopped after 5 weeks, so the fruits could not be described. However, in a previous paper, the slow growth and delay of flowering were not mentioned as characteristics of this line. It was described as being identical to the wild type, except for the fruits that had a lighter tint and were smooth (Skarzyńska *et al.*, 2017).

As for S3 line, in this study it had a slower growth and might even be a dwarf plant as it wass never greater than the half height of the wild type B10. It also had a different shape of leaves and the whole plant had a lighter tint. Similarly to the S2 line, it was described differently in a previous paper. It was described as being identical to the wild type except for the lighter tint of the whole plant (Skarzyńska *et al.*, 2017).

The noticeable differences between this study and the publication from 2017 can be explained because the plants are from the next generation. Indeed, the changes that are already present in the genotype could be perpetuated from one generation to the other and they could be more visible in the phenotype.

IV.2. Statistics on differentially expressed genes

In this study 364 DEGs were identified in S2 and 273 in S3 through RNA-seq. In 2016, a publication on sweet cherry identified 39 different proteins only between the 'Hedelfinger' cultivar and its somaclonal variant, at two ripening stages (Prinsi *et al.*, 2016). Another study using AFLP analysis of cDNA showed only 62 DEGs in sugarcane somaclones in response to infection with *Ustilago scitaminea* or *Bipolaris sacchari* compared to non-resistant sugarcane plants (Borrás-Hidalgo *et al.*, 2005). These publications suggest that the amount of DEGs found in the cucumber somaclones of this study is particularly high, and thus, hard to study completely in so few pages. Moreover, both articles continue their analysis with the relation between the function of the DEGs and the phenotype. Nevertheless, such studies as chromosome maps, heatmaps and promoter analysis are usually not combined with this kind of analysis for somaclones characterization. However, most of the studies of somaclone characterization rely on different markers (mostly RAPD markers) to study the genetic variability of somaclones (Leva *et al.*, 2012; Matheka *et al.*, 2008; Martín *et al.*, 2002; Damasco *et al.*, 1998).

IV.3. Verification of RNA-seq results by qPCR line

RNA-seq data were confirmed at 100% for S2 line and at 94% for S3 line. In other articles, rates of confirmed genes were about 86-96% and so were similar to the ones found in this study (Pawełkowicz *et al.*, 2019; O'Rourke *et al.*, 2013; Gao *et al.*, 2013). The differences obtained between the two techniques can be explained by their difference of sensitivity (Pawełkowicz *et al.*, 2019). Only one gene was not confirmed due to the non-significance of the result, and despite the consistent trend of the gene expression in qPCR with the RNA-seq data. This gene is *Cucsat.PASA.G149* and corresponds in *Arabidopsis thaliana* to the gene

TPS21, also called *AT5G23960*. It encodes for a sesquiterpene synthase that is responsible with only one other sesquiterpene synthase to the production of all the group of sesquiterpenes found in the *Arabidopsis thaliana* floral volatile blend (Tholl *et al.*, 2005). These molecules might play a role in attracting insects to the plant even if *Arabidopsis* emits fewer floral volatiles than the other insect-pollinated plants, because it is mostly self-pollinated (Chen *et al.*, 2003). The difference of expression between the wild type B10 and the somaclonal line S3 for this gene is not significant. The non-significance of the result can be explained by differences found between the biological replicates that were used to calculate the standard deviation for all genes and all lines due to their position in the field. Indeed, it is possible that the plants are not all under exactly the same conditions. Indeed, in the promoters' analysis, it is clearly visible that there are a substantial number of light response elements in all the analyzed promoters' sequences, and so there were probably some plants that were more in the shadow than others in the field. Moreover, *Cucsat.PASA.G149* possesses a lot of light responsive elements in his promoter region. The different conditions of lighting between the biological replicates of this gene is probably the reason why the standard deviation is that high, making the result non-significant. However, to the knowledge of the author, no study on the impact of heterogeneity between replicates would have been carried out until this day.

IV.4. Analysis of protein interaction through bioinformatics tools

Several groups were determined in the protein networks of both somaclonal lines. In the first group, there are proteins related to the cell cycle. It is the basis of cellular multiplication and therefore is an important element in plant growth. In plants, it is known to be regulated by a large number of TF (Dewitte and Murray, 2003). However, the TF that are differentially expressed in the somaclonal lines do not correspond to the specific TF acting at the level of the cell cycle, so it cannot explain why some of the proteins implied in the cell cycle are differentially expressed. Nevertheless, it is also known that the cell cycle is highly dependent on hormone activity like auxins, cytokinins, brassinosteroids, abscisic acid and jasmonic acid (Dewitte and Murray, 2003). In S3 line, some auxin-related genes are differentially expressed while in S2 line, there are DEGs related to all hormones quoted before except the jasmonic acid. It implies a disrupted hormones expression and might explain the changes in the cell cycle. These changes are confirmed by the phenotypical analysis of the somaclonal lines that both show a slower growth or even a potential dwarfism for S3 line. But surprisingly, S3 does not show any DEG related to dwarfism while S2 possesses one.

The second group concerned the proteins implied in transcription. There are some proteins that are directly responsible for the transcription like the subunits of the RNA polymerase while the other proteins are TF in both somaclonal lines. The transcription process is an important process of a cell life since it enables the expression of all the genes that can be found in the DNA of a living being. However, all the genes are not constitutively expressed. They are all regulated by TF that are themselves acting depending on different factors that can be intracellular or extra cellular, including the plant environment. There are TF for every developmental process. The PLATZ TF are known to be implied in cell division and can act as a repressor or as an activator (Kim *et al.*, 2018). The HD-Zip TF are mostly involved in plant development (Ariel *et al.*, 2007). Their presence can partly explain the slow growth of S2 and S3 for PLATZ, and only S2 for the HD-Zip. The TF containing a NAC domain are mostly

known to be implied in biotic and abiotic stress tolerance (Puranik *et al.*, 2012). It was shown that WRKY can be differentially expressed in response to wounding (Eulgem *et al.* 2000), what happened to the plants in both somaclonal lines. The bHLH TF have a very wide range of actions, including stress response (Feller *et al.*, 2010). The ERF TF are mostly regulating stress response (Xu *et al.*, 2008). The Heat Shock TF are implied in the expression of the heat shock proteins under stress conditions (Lohmann *et al.*, 2004). The 3 first ones are present in S2 and S3 and the 2 latter are present only in S2. They might be the reflection of the stressful conditions that the plants underwent during *in vitro* culture. The Heat Shock TF might bind to a stress responsive element found in the promoter region of S2 DEGs. Moreover, the bHLH TF is also implied in hormone signaling, like DOF and bZIP TF (Feller *et al.*, 2010; Noguero *et al.*, 2013; Jakoby *et al.*, 2002) and the KAN TF is known to regulate auxin transport (Ilegems *et al.*, 2010). In both somaclonal lines, hormone responsive elements were important *cis*-acting elements in the promoter region. Hence, the 3 first TF might bind to some of them. The KAN TF might imply changes in the hormonal balance of S2 line and so a disrupted growth and flowering, partly explaining the features of this line. Nevertheless, there is a DOF TF only in S3 line. Besides, bHLH and MYB TF can interact together in the flavonoid biosynthesis for instance (Feller *et al.*, 2010). That might explain the presence of a few flavonoid biosynthetic genes regulation in the promoter region of S2 and S3 DEGs. Nonetheless, one of the MYB TF detected in S2 line is acting in response to UV-B light. DOF and bZIP TF are also known to be implied in light signaling (Noguero *et al.*, 2013; Jakoby *et al.*, 2002). The TF with a GATA domain are often depending on light (Reyes *et al.*, 2004). The MYB TF might be regulated by a light responsive element in S2, while the last ones would bind to this kind of light responsive element in S3 for DOF, S2 for GATA and in both lines for bZIP. The changes in light perception and hormone balance in the plant have an effect on flowering. Indeed, the bZIP and bHLH TF can also be implied in flower development as well as MADS-box and AP2 TF that are really numerous and C_2H_2 TF (Jakoby *et al.*, 2002; Feller *et al.*, 2010; Riechmann and Ratcliffe, 2000). The TF with a jmjC domain is known to have an epigenetic impact by demethylation of DNA. The proteins of this family mostly regulate regions implied in flowering or in the circadian cycle (Chen *et al.*, 2011). This type of TF can partly explain the delay in flowering in S2 and S3 for jmjC, bZIP, bHLH and S2 only for MADS-box, AP2 and C_2H_2. The bHLH TF is also implied in phytochrome signaling (Feller *et al.*, 2010). It might explain the presence of phytochrome expression elements in the promoter region of both lines.

The 3rd group concerned the proteins involved in the translation process. It is a process as important as transcription since this is the next step that is necessary in the plants' genes expression. Translation enables to build proteins from a specific mRNA. The DEGs of S3 line encode for the protein responsible for the elongation of the protein. It can be concluded that all the translation process should be impacted, for all genes. However, it is impossible to confirm such hypothesis with the available information obtained during this study.

The 4th group of proteins was related to flowering processes. Flowering is a very intricate part of a plant's life, regulated by a very wide range of genes. As for the TF implied in flowering, these proteins in S3 line might distort the natural flowering process of cucumber. This implies some changes in the flowering and it can be a delay in the first appearance of the flowers, explaining the phenotype analysis of S2 and S3 lines.

In the 5th group were classified the transporters. They are essential molecules in any living being as they enable the communication inside the cell, between all the organelles, or

even between the cells themselves. Without transporters, the coordination of all the cells would not be possible, and so complex pluricellular organism could not exist either. Most of the transporters detected in S3 line are ion transporters, especially Ca^{2+} transporters. This ion is one of the most important second messenger in plants. Its action can be regulated in response to hormones, light or stress. Moreover, some enzyme cannot work without Ca^{2+}, for instance enzymes involved in stress tolerance. It is involved in cell cycle regulation as well (Tuteja and Mahajan, 2007). Thus, if the transport of such an important ion is disrupted, it can be expected to see important changes in the plants, like a differential expression of some hormone, light or stress responsive genes. The promoters' analysis can confirm this hypothesis since hormone, light and stress responsive elements were among the most important elements in the promoter region of S3 line. Besides, it can be suggested that the Ca^{2}+ transporters reflect the stress underwent by the plants during *in vitro* culture.

The 6th group was dedicated to the peroxisome metabolism. Photorespiration takes place in this organelle but a lot of different processes happen there as well, like events of the secondary metabolism, some developmental processes and part of the stress response of the plant (Hu *et al.*, 2012). It is difficult to judge how the differential expression of these proteins could impact the primary and secondary metabolism of the plant as the research was not oriented on these elements, but it can show once again the reflection of the stress of the *in vitro* culture at a moment of the plants life.

The 7th and last group was attributed to detoxification processes. The proteins of this group are mostly glutathione S-transferases or other peroxidases that are both known to be implied in detoxification processes by removing the toxic hydroperoxides from the plants' cells (Edwards *et al.*, 2000). The hydroperoxides like H_2O_2 are Reactive Oxygen Species (ROS) that have an important role of signaling in plants. They are implied in processes of growth and development but also in apoptosis and in response to stresses. However, these molecules can be very harmful for the organism (del Río, 2015). That is why the excess needs to be removed by antioxidative molecules like the proteins of this group which were found in S3 line. The differential expression of the detoxification proteins can once again be based on the stress provoked by *in vitro* culture, creating a lot of ROS and thus a need to remove their excess.

No further analysis of the S2 network will be done as it is very complex. However, it could be the subject of a whole study in a later work. It would have been interesting to make a more intricate analysis drawing a parallel between the STRING analysis and the qPCR and RNA-seq results, to determine how big can be the impact of each studied protein in the plant.

IV.5. Bioinformatics analysis of promoter regions of DEGs in somaclonal lines

PlantCare analysis showed that 33% of the motifs found in the promoters' sequences in S2 and 31% in S3 are core promoter elements. The core promoter elements are necessary to the transcription because they include the starting site of transcription. They can be made up of motifs like TATA-box, which is one of the most important ones (Burke and Kadonaga, 1997). The CAAT-boxes are also counted in the core promoter elements (Lee-Huang *et al.*, 1993) and they have an influence on the transcriptional initiation frequency (Kusnetsov *et al.*, 1999). This explains their abundance among the analyzed promoter sequences, since the genes cannot be

transcribed and so expressed in any way without the presence of a core promoter element. The other motifs, less abundant in the analyzed sequences are more specific to the genes. Most of them regulate gene expression as an answer to different factors such as light, stress or hormones. The motif MYB is known to regulate genes that are mostly specialized in cell proliferation and cell shape, in response to hormones like gibberellic and abscisic acids (Martin and Paz-Ares, 1997). The motif MYC regulates genes that might be involved in cell cycle, differentiation and death (Amati and Land, 1994). It is also implied in some hormones' pathways like the pathway of jasmonic acic in response to a stress (Lorenzo *et al.*, 2004; Boter *et al.*, 2004). MYC genes are interacting with MYB genes (Lorenzo *et al.*, 2004). This can explain why both motifs are found in approximately equivalent quantities among the analyzed sequences in both lines. As for the motif ABRE, it is implied in stress tolerance like drought tolerance in response to ABA (Yoshida *et al.*, 2010). The CGTCA-motif and TCACG-motif of S3 line are both related to stress in response to methyl jasmonate (Sirhindi *et al.*, 2016). Other motifs like G-boxes can regulate multiple genes depending on a wide range of factors. They are not specific but are not in every promoter region or necessary to the translation (Menkens *et al.*, 1995). Their wide range of possible actions explains their quite substantial presence in the studied sequences.

The short motifs between 4 and 6 nucleotides were highly represented among the analyzed promoters' sequences while the longer ones were way less present. Indeed, the TATA-boxes and CAAT-boxes are very small motifs of 4 or 5 nucleotides while the other more specific motifs have longer sequences. That explains the distribution of the lengths of the motifs found in the results.

The functions of the motifs are probably the most important characteristic to study in the promoter region analysis as it determines under which conditions the gene is regulated. As seen previously, a lot of them are acting in response to hormones. They represent a quite weak percentage of the analyzed sequences but remain the fourth and fifth more represented category of function in S2 and S3 lines respectively. The stress response function is of the same rough estimate, as it is the fifth most represented category in S2 line and the fourth in S3 line. Just after them come the oxygen responsive elements in which elements related to anoxia are classified. This abundance of hormone and stress responsive elements suggests that the plants were facing stress and endured some changes in their hormonal response and/or environment. Plant hormones have a very wide range of different effects in the plants. They are regulating growth and development, including the cell growth, division and differentiation and its adaptation to the plant environment. They regulate also the flowering time and the following fruit development, ripening and even fall, as well as the leaves' fall. They regulate seed germination. They have also other roles outside the growth and development aspect of every organ, like the control of the stomatal opening or the regulation of stress response. In a few words, hormones regulate almost every aspect of the plant life from the seed to its death (Davies, 1995). Knowing that, it becomes clear that the smallest change in the plant environment can induce some changes in its hormonal balance. Moreover, the plants used in this study went through a lot of changes. Indeed, *in vitro* culture needs to put the plant in a stress state, because an explant is cut from the whole plant. Then this explant is sterilized and put in a medium totally different from its previous environment, in which anoxia can sometimes occur, like in the liquid medium where S3 line was grown. The wound and the products used for sterilization as well as the lack of natural molecules provided by the plant where it came from, put the explant in a great stress situation, implying the activation of stress responsive genes. Furthermore, as it was said previously, hormones are implied in the regulation of stress

response. It can be concluded that such a stress changes the hormone balance in the explant and so changes the expression of hormone responsive genes. Another reason for the change in the hormonal balance of the explant can be explained by the composition of the medium itself. Often, some substitute artificial hormones are used to replace the natural hormones that were previously provided by the plant. . The most common used medium in *in vitro* culture nowadays is the MS medium (Musharige and Skoog, 1962) elaborated by Musharige and Skoog in the sixties (Thorpe, 2007). The synthetic hormones added to this medium have different aims. Synthetic auxins, for instance, are used for callus induction or for the organogenesis of roots from a callus. Auxin enables the root primordia formation but inhibits its growth while gibberellins inhibits the root primordia formation but is necessary for its growth. As for cytokinins, they are used for the development of adventitious buds and for cell multiplication in combination with auxins (Gaspar *et al.*, 1996). Moreover, the somaclonal lines of this study were created on MS medium supplemented with hormones: 2,4-D (an auxin) and IPA (a cytokinin) for S2 line that regenerated passing through a callus stage and for S3, 2,4-D was used to obtain a callus that was then put in a solution with BAP (a cytokinin) to regenerate plants from single callus cells. All these facts are enough to explain a change in the genes expression but not why they are still present in a fully-grown plant that went through a field culture after the *in vitro* stage. However, the plants stayed a long time under stressful conditions (4 to 10 months) and it is well known that a long exposure to stress can induce mutations or epigenetic changes for the plant to adapt to its environment. It was proven through different works that such kind of changes can be induced under plant tissue culture (Phillips *et al.*, 1994; Kaeppler *et al.*, 2000). Moreover, already in 1990, it was said that hormones are considered as potential mutagenic factors, as well as the stress of *in vitro* culture, anoxia and the lack of elements previously provided by the plant (Tonelli, 1990). It is then possible that these changes occurred in the genes with hormone responsive elements and in the genes with stress responsive elements in their promoter region, as well as in the genes with oxygen responsive element because of potential anoxia conditions. This might explain how the changes of expression that may have been observed during *in vitro* culture of B10 line were preserved through the different stages of the culture, creating new variants and explaining the difference of expression of the DEGs between B10 line and both somaclonal lines. A similar case was studied on the oil palm where the somaclonal variants showed a difference of gene expression compared to the wild type in normal conditions but also in response to auxin (Morcillo *et al.*, 2006). In addition, as said previously, the changes in the hormone responsive genes affect organs growth and development, so it is not surprising to find seed-specific regulation elements in the promoter region of S2 DEGs and organ-specific regulation elements (it includes root-specific and seed-specific regulation elements) in the promoter region of S3 DEGs. Hormones are also regulating cell differentiation (Davies, 1995) and a disrupted hormonal regulation might be the reason of the presence of cell differentiation elements in the promoter region of both somaclonal lines' DEGs.

The most abundant category of function was the light responsive elements. Plant's life, in most of the cases, is conditioned by light as it is necessary for the photosynthesis that enables the production of sugars for plant growth and development. Some artificial light was used during *in vitro* culture. It is only rarely that natural light is used for plant tissue culture, like in the experiment of Kodym and Zapata-Arias on banana in 1998 where they qualify the natural light as an alternative light source. These lights are not of the same intensity and with the same wavelength than the sun light. The photoperiod is every day the same, usually 16 hours of light

and 8 hours of night (Kodym and Zapata-Arias, 1998; Comino *et al.*, 2019), and the light does not turn on and off gradually as the sun rises or sets. These artificial conditions imply a different expression of the genes that are regulated by light responsive elements and a long exposure to unusual lighting conditions can cause a long-term adaptation of the plant implying mutations and epigenetic changes. Indeed, it was found in another study that a somaclone of *Prunus avium* L. cv. Hedelfinger showed a different response to the light compared to the wild type (Piagnani *et al.*, 2002). It is likely that the same kind of mutation touching the light responsive genes occurred in S2 and S3 line. Additionally, a modified lighting and/or response to light affects the circadian cycle of plants (McWatters *et al.*, 2000) and might explain the presence of circadian cycle regulation elements in the analyzed sequences. Besides, these changes can modify the expression of different plant photoreceptors like the phytochromes and cryptochromes (Tóth *et al.*, 2001), giving an explanation to the presence of phytochrome expression elements, even in small quantities, in the analyzed sequences. Moreover, photosynthesis is highly dependent of the temperature and thus it can be expected to see a strong link between light responsive genes and temperature responsive genes (Long *et al.*, 1994). This dependence might be an interesting element to explain the presence of temperature responsive elements in the analyzed sequences. Flowering is another characteristic that can be affected by light. Cucumber flowers are not all induced under the same conditions: male flowers are induced under long days and female flowers are induced under short days (INFOARGO SYSTEMS, SL, 2014). The cultivar used in this study, the 'Borszczagowski' cultivar, gives only male flowers. Besides, if the circadian cycle of the plant is disrupted, as it might be in the somaclonal lines, the flowering time can be affected, as seen in the phenotype of both lines. However, what is quite surprising is that only 1 flowering regulation element was detected in the promoter region of S2 DEGs, while the delay of flowering is visible in both somaclonal lines.

The last five functions, zein metabolism regulation elements, meristem expression elements, endosperm expression elements, flavonoid biosynthetic genes regulation and cell cycle regulation elements, might be implied in disrupted developmental processes (Zhou *et al.*, 2017). It can be explained by the modified hormone expression in somaclonal lines since hormones regulate the development of plants. Moreover, these disrupted developmental processes can be seen in the phenotype of the somaclonal lines. Indeed, S2 shows a slow growth and a delay in flowering while S3 has not only a slow growth but might also be a dwarf variant because it never exceeds the half height of the wild type B10 in 5 weeks of culture, and it also shows a delay in flowering, as seen in the phenotypical analysis of the somaclonal lines.

IV.6. Chromosomal location of DEGs

The DEGs of both somaclonal lines were mapped on their corresponding cucumber chromosomes and contigs. The chromosomal location of the DEGs is not evenly distributed. Moreover, the number of DEGs mapped on a chromosome does not depend on its length. From this analysis, it can be suggested that the DEGs are not totally random and that there are links between at least some of them. Another study on HeLa cells and human primary fibroblasts points out that the non-coding sequences are usually randomly distributed on chromosomes while the place of a gene is highly characteristic for this gene although independent from its expression (Kurz *et al.*, 1996). Moreover, a master's thesis on a cucumber somaclone concluded

that the identified DEGs are not randomly distributed on the chromosomes (Mróz, 2019). It is probable that the differential expression of one gene impacted the expression of other close genes and it would explain the more or less large groups DEGs mapped together. This hypothesis is confirmed by the STRING analysis where the PPI enrichment p-value is significant, which means that the DEGs put in the analysis are not randomly chosen genes and that they belong to groups, as it can be seen on the chromosomes' maps. However, a further analysis could be made to check if the genes implied in networks correspond to the groups of genes found on the chromosomes or if they are totally different. Of course, all the genes of the main network do not belong to the same groups on the chromosomes' maps because there is a greater number of DEGs in this network than in any of these groups, but it might be true for some of them. It also means that some genes located on different chromosomes or far from each other can also be somehow connected and are not necessarily random. Nevertheless, only 84% of contigs were mapped to the chromosomes (Osipowski *et al.*, under review) so the hypothetical length of the chromosomes compared to each other in this study might be totally wrong. Indeed, in S2 line, there are 15 contigs that are not mapped on chromosomes (they bear some of the DEGs studied here) and in S3 line, there are 46 contigs that are not mapped. It means that in total, the chromosomes' maps elaborated in this study miss at least 46 contigs, if not more. Hence, it is likely that any other chromosome could be longer than the 3rd that is considered as the longest, even the 7th one that is considered to be the shortest one with the currently possessed information. Consequently, this analysis is only valid in the hypothesis that the missing contigs are evenly divided between each chromosome.

IV.7. Heatmap of DEGs

The heatmaps show that most of the DEGs have a consistent trend of expression between their three biological replicates. Only very few of them are not consistent. It can be concluded that the environment had a minimal impact on the replicates growth and that the experiment is reproducible. Moreover, the HCL analysis showed in both somaclonal lines 2 anti-correlated groups of DEGs having inverted trend of expression. Inside each group all the genes have a rather correlated expression. From this analysis it can be concluded that the DEGs belong to 2 distinct groups which expression trend is inverted. However, the medium Pearson correlation coefficient points out that the trend of expression of the DEGs in a group is not random, which confirms the previous analyses.

To conclude this discussion, it can be noticed that S2 and S3 are really similar through most of the analyses. The biggest differences lie in the number of identified DEGs which is much higher in S2 than in S3, their nature and mostly their connections. These differences are probably the explanation for their phenotypical differences in the color and supposedly in the fruits. However, they also have common points in the phenotype (delay in growth and flowering) that were previously explained through the analysis of the category of genes they have in common. Moreover, as epigenetic changes were detected, it could be interesting to develop more widely this aspect and link it to phenotypical changes as it was done in a study of wheat somaclones (Baránek *et al.*, 2016).

V. CONCLUSION

The somaclonal lines S2 and S3 both show phenotypical changes compared to the wild type line B10. S2 seems morphologically similar to B10 during the 5 weeks of culture when the plants were measured. However, a delay of growth was noticeable during the 4 first weeks of culture as well as a delay in flowering visible at the 5th week of culture, and in a previous description of the line, it was said that its fruit had a lighter color and was smooth, unlike B10 fruits. As for S3 line, it had an even stronger delay of growth, without flowering during the 5 weeks of culture. It might be a dwarf variant. It also shows a different shape of the leaves and a lighter tint of the whole plant.

To explain these features, it was necessary to check the repeatability of the experiment and the correctness of the RNAseq data. The genes were confirmed at 100% for S2 line and 94% for S3 line because of a lack of accuracy that can be explained by the difference of sensitivity between the methods. Thus, it can be considered that RNAseq results are reliable. Moreover, the heatmap showed that most of the DEGs have a consistent trend between the 3 biological replicates. Therefore, the experiment can be considered repeatable. Besides, the gene map shows that the DEGs are not random and so, that they are linked together and can be categorized in groups, giving meaning to the executed analysis.

The analysis of the protein network and of the promoter region of the DEGs in both somaclonal lines enabled to explain some of their features. The slow growth can for instance be explained by the DEGs implied in the cell cycle, and by the ones related to hormones as well as some transcription factors like PLATZ and HD-Zip. Indeed, these transcription factors can be under the regulation of hormone responsive promoter elements. Besides, proteins implied in cell cycle can potentially be under the control of a transcription factor implied in cell division, explaining their differential expression. Moreover, a lot of promoter elements could be responsible for these developmental changes like the zein metabolism regulation elements, meristem expression elements, endosperm expression elements, flavonoid biosynthetic genes regulation and cell cycle regulation elements that are known to be implied in developmental processes. The different shape of S3 leaves can also be explained by these factors. However, S3 is not confirmed as being a dwarf variant because no DEGs was found corresponding to this state in this line but one was curiously found in S2.

Flowering is regulated by a lot of different factors. The one mostly found in both somaclonal lines was the DEGs related to hormones, potentially explaining the late flowering of both somaclonal lines. Some transcription factors might be implied directly in flowering or through their relation to hormones, light and/or circadian cycle. These 2 latter change the perception of the photoperiod, another factor implied in flowering.

Nonetheless, no explanation was found to explain the supposed difference of fruits in S2 or the different color of the whole plant in S3.

The network of S2 was very hard to divide into groups as the main network counts 2 main central nodes of very different functions that are linked to a great variety of proteins, often impossible to put in the same group because of their differences. Because of this issue, the analysis was not performed with as much accuracy as it should be. In further researches, each protein of the networks should be compared with the RNAseq analysis to determine more precisely their role in the phenotype and maybe find the explanation of the features that were not justified in this work.

BIBLIOGRAPHY

Acquaah G., Clonal propagation and *in vitro* culture, In: Acquaah G. (ed), Principles of plant genetics and breeding, Second edition, 2012, p.146-170, DOI 10.1002/9781118313718.ch8

Aimin L., Linbao X., Yongtai Z., Feihu H., 2004, Research Progress of Tactics for Cucumber Mosaic Virus Control, In: *Journal of Changjiang Vegetables*

Amati B., Land H., 1994, Myc—Max—Mad: a transcription factor network controlling cell cycle progression, differentiation and death, In: *Current Opinion in Genetics & Development*, Vol. 4, Issue 1, p.102-108, DOI 10.1016/0959-437X(94)90098-1

Ando K. and Grumet R., 2010, Transcriptional Profiling of Rapidly Growing Cucumber Fruit by 454-Pyrosequencing Analysis, In: *Journal of the American Society for Horticultural Science*, Vol. 135, Issue 4, p.291-302, DOI 10.21273/JASHS.135.4.291

Ariel F.D., Manavella P.A., Dezar C.A., Chan R.L., 2007, The true story of the HD-Zip family, In: *Trends in Plant Science*, Vol. 12, Issue 9, p.419-426, DOI 10.1016/j.tplants.2007.08.003

Baránek M., Čechová J., Kovacs T., Eichmeier A., Wang S., Raddová J., Nečas T., Ye X., 2016, Use of Combined MSAP and NGS Techniques to Identify Differentially Methylated Regions in Somaclones: A Case Study of Two Stable Somatic Wheat Mutants, In: PLoS ONE, Vol. 11, n°10, 21 p., DOI 10.1371/journal.pone.0165749

Bhatia S., Sharma K., Technical glitches in micropropagation, In: Bhatia S., Sharma K., Dahiya R., Bera T. (eds), Modern applications of plant biotechnology in pharmaceutical sciences, 2015, p.393-404, DOI 10.1016/B978-0-12-802221-4.00013-3

Borrás-Hidalgo O., Thomma B.P.H.J., Carmona E., Borroto C.J., Pujol M., Arencibia A., Lopez J., 2005, Identification of sugarcane genes induced in disease-resistant somaclones upon inoculation with *Ustilago scitaminea* or *Bipolaris sacchari*, In: *Plant Physiology and Biochemistry*, Vol. 43, Issue 12, p.1115-1121, DOI 10.1016/j.plaphy.2005.07.011

Boter M., Ruíz-Rivero O., Abdeen A., Prat S., 2004, Conserved MYC transcription factors play a key role in jasmonate signaling both in tomato and Arabidopsis, In: *Genes & Development*, Vol. 18, p.1577-1591, DOI 10.1101/gad.297704

Burke T.W., Kadonaga J.T., 1997, The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*, In: *Gene & Development*, Vol. 11, p.3020-3031, DOI 10.1101/gad.11.22.3020

Burza W., Malepszy S., 1995, Direct plant regeneration from leaf explants in cucumber (*Cucumis sativus* L.) is free of stable genetic variation, In: *Plant Breeding*, Vol. 114, Issue 4, p.341-345, DOI 10.1111/j.1439-0523.1995.tb01246.x

Cassells A.C., Curry R.F., 2001, Oxidative stress and physiological, epigenetic and genetic variability in plant tissue culture: implications for micropropagators and genetic engineers, In: *Plant Cell, Tissue and Organ Culture*, Vol. 64, Issue 2-3, p.145-157, DOI 10.1023/A:1010692104861

Chandrasekaran J., Brumin M., Wolf D., Leibman D., Klap C., Pearlsman M., Sherman A., Arazi T., Gal-On A., 2016, Development of broad virus resistance in non-transgenic cucumber

using CRISPR/Cas9 technology, In: *Molecular Plant Pathology*, Vol. 17, Issue 7, p.1140-1153, DOI 10.1111/mpp.12375

Chen C., Liu M., Jiang L., Liu X., Zhao J., Yan S., Yang S., Ren H., Liu R., Zhang X, 2014, Transcriptome profiling reveals roles of meristem regulators and polarity genes during fruit trichome development in cucumber (*Cucumis sativus* L.), In: *Journal of Experimental Botany*, Vol. 65, Issue 17, p.4943-4958, DOI 10.1093/jxb/eru258

Chen F., Tholl D., D'Auria J.C., Farooq A., Pichersky E., Gershenzon J., 2003, Biosynthesis and Emission of Terpenoid Volatiles from Arabidopsis Flowers, In: *The Plant Cell*, Vol. 15, Issue 2, p.481-494, DOI 10.1105/tpc.007989

Chen X., Hu Y., Zhou D.-X., 2011, Epigenetic gene regulation by plant Jumonji group of histone demethylase, In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, Vol. 1809, Issue 8, p.421-426, DOI 10.1016/j.bbagrm.2011.03.004

Chen Z.-F., Kang X.-P., Nie H.-M., Zheng S.-W., Zhang T.-L., Zhou D., Xing G.-M., Sun S., 2019, Introduction of Exogenous Glycolate Catabolic Pathway Can Strongly Enhances Photosynthesis and Biomass Yield of Cucumber Grown in a Low-CO2 Environment, In: *Frontiers in Plant Science*, Vol. 10, Article 702, 11 p., DOI 10.3389/fpls.2019.00702

Comino C., Moglia A., Repetto A., Tavazza R., Globe Artichoke Tissue Culture and Its Biotechnological Application, In: Portis E., Acquadro A., Lanteri S. (eds), The Globe Artichoke Genome. Compendium of Plant Genomes, 2019, p.41-64, DOI 10.1007/978-3-030-20012-1_3

Custers J.B.M., Bergervoet J.H.W., Somaclonal variation as a means of overcoming post-fertilization barriers in interspecific crosses in Cucumis and Lactuca, In: Novak F.J., Havel L., Dolezel J. (eds), Plant tissue and cell culture-application to crop improvement, 1984, p.509-510

Custers J.B.M., Ziljstra S., Jansen J., 1990, Somaclonal variation in cucumber (*Cucumis sativus* L.) plants regenerated via embryogenesis, In: *Acta Botanica Neerlandica*, Vol. 39, n°2, p.153-161, DOI 10.1111/j.1438-8677.1990.tb01483.x

Damasco O.P., Smith M.K., Adkins S.W., Hetherington S.E., Godwin I.D., 1998, Identification and characterisation of dwarf off-types from micropropagated Cavendish bananas, In: Acta Horticulturae, Vol. 490, p.79-84, DOI 10.17660/ActaHortic.1998.490.5

Davies P.J., The Plant Hormones: Their Nature, Occurrence and Functions, In: Davies P.J. (ed), Plant Hormones: Physiology, Biochemistry and Molecular Biology, 1995, p.1-12, DOI 10.1007/978-94-011-0473-9

Dewitte W., Murray J.A.H., 2003, The Plant Cell Cycle, In: *Annual Review of Plant Biology*, Vol. 54, n°1, p.235-264, DOI 10.1146/annurev.arplant.54.031902.134836

Edwards R., Dixon D.P., Walbot V., 2000, Plant glutathione S-transferases: enzymes with multiple functions in sickness and in health, In: *Trends in Plant Science*, Vol. 5, Issue 5, p.193-198, DOI 10.1016/S1360-1385(00)01601-0

Eulgem T., Rushton P.J., Robatzek S., Somssich I.E., 2000, The WRKY superfamily of plant transcription factors, In: *Trends in Plant Science*, Vol. 5, Issue 5, p.199-206, DOI 10.1016/S1360-1385(00)01600-9

Feller A., Machemer K., Braun E.L., Grotewold E., 2010, Evolutionary and comparative analysis of MYB and bHLH plant transcription factors, In: *The Plant Journal*, Vol. 66, Issue 1, Special Issue: The Plant Genome: An Evolutionary View on Structure and Function, p.94-116, DOI 10.1111/j.1365-313X.2010.04459.x

Gao L., Jin Tu Z., Millett B.P., Bradeen J.M., 2013, Insights into organ-specific pathogen defense responses in plants: RNA-seq analysis of potato tuber-*Phytophthora infestans* interactions, In: *BMC Genomics*, Vol. 14, Article 340, 12 p., DOI 10.1186/1471-2164-14-340

Gaspar T., Kevers C., Penel C., Greppin H., Reid D.M., Thorpe T.A., 1996, Plant hormones and plant growth regulators in plant tissue culture, In: In Vitro *Cellular & Developmental Biology – Plant*, Vol. 32, Issue 4, p.272-289, DOI 10.1007/BF02822700

Götz S., García-Gómez J.M., Terol J., Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talón M., Dopazo J., Conesa A., 2008, High-throughput functional annotation and data mining with the Blast2GO suite, In: *Nucleic Acids Research*, Vol. 36, Issue 10, p.3420-3435, DOI 10.1093/nar/gkn176

Guo S., Zheng Y., Joung J.-G., Liu S., Zhang Z., Crasta O.R., Sobral B.W., Xu Y., Huang S., Fei Z., 2010, Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types, In: *BMC Genomics*, Vol. 11, Article 384, 13 p., DOI 10.1186/1471-2164-11-384

Guo W.L., Wu R., Zhang Y.F., Liu X.M., Wang H.Y., Gong L., Zhang Z.H., Liu B., 2007, Tissue culture-induced locus-specific alteration in DNA methylation and its correlation with genetic variation in Codonopsis lanceolata Benth. et Hook. f., In: *Plant Cell Reports*, Vol. 26, n°8, p.1297-1307, DOI 10.1007/s00299-007-0320-0

Hao Y.-J., Deng X.-X., 2002, Occurrence of chromosomal variations and plant regeneration from long-term-cultured citrus callus, In: *In Vitro Cellular & Developmental Biology – Plant*, Vol. 38, Issue 5, p.472-476, DOI 10.1079/IVP2002317

Hao Z., Fan C., Cheng T., Su Y., Wie Q., Li G., 2015, Genome-Wide Identification, Characterization and Evolutionary Analysis of Long Intergenic Noncoding RNAs in Cucumber, In: *PLoS ONE*, Vol. 10, n°3, 20 p., DOI 10.1371/journal.pone.0121800

Hellemans J., Mortier G., De Paepe A., Speleman F., Vandesompele J., 2007, qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, In: *Genome biology*, Vol. 8, Issue 2, Article R19, DOI 10.1186/gb-2007-8-2-r19

Hu B., Li D., Liu X., Qi J., Gao D., Zhao S., Huang S., Sun J., Yang L., 2017, Engineering Non-transgenic Gynoecious Cucumber Using an Improved Transformation Protocol and Optimized CRISPR/Cas9 System, In: *Molecular Plant*, Vol. 10, p.1575-1578, DOI 10.1016/j.molp.2017.09.005

Hu J., Baker A., Bartel B., Linka N., Mullen R.T., Reumann S., Zolman B.K., 2012, Plant Peroxisomes: Biogenesis and Function, In: *The Plant Cell*, Vol. 24, Issue 6, p.2279-2303, DOI 10.1105/tpc.112.096586

Ilegems M., Douet V., Meylan-Bettex M., Uyttewaal M., Brand L., Bowman J.L., Stieger P.A., 2010, Interplay of auxin, KANADI and Class III HD-ZIP transcription factors in vascular tissue formation, In: *Development*, Vol 137, n°6, p.975-984, DOI 10.1242/dev.047662

Jain S.M., 2001, Tissue culture-derived variation in crop improvement, In: *Euphytica*, Vol. 118, Issue 2, p.153-166, DOI 10.1023/A:1004124519479

Jakoby M., Weisshaar B., Dröge-Laser W, Vicente-Carbajosa J., Tiedemann J., Kroj T., Parcy F., 2002, bZIP transcription factors in *Arabidopsis*, In: *Trends in Plant Science*, Vol. 7, Issue 3, p.106-111, DOI 10.1016/S1360-1385(01)02223-3

Jevremović S., Subotić A., Miljković D., Trifunović M., Petrić M., Cingel A., 2012, Clonal fidelity of chrysanthemum cultivars after long term micropropagation by stem segment culture, In: *Acta Horticulturae*, Vol. 961: VII International Symposium on In Vitro Culture and Horticultural Breeding, p.211-216

Kaeppler S.M., Kaeppler H.F., Rhee Y., Epigenetic aspects of somaclonal variation in plants, In: Matzke M.A., Matzke A.J.M. (eds), Plant Gene Silencing, 2000, p.59-68, DOI 10.1007/978-94-011-4183-3_4

Khan S, Saeed B., Kauser N., 2011, Establishment of genetic fidelity of *in-vitro* raised banana plantlets, In: *Pakistan Journal of Botany*, Vol. 43, n°1, p.233-242

Kim J.H., Kim J., Jun S.E., Park S., Timilsina R., Kwon D.S., Kim Y., Park S.-J., Hwang J.Y., Nam H.G., Kim G.-T., Woo H.R., 2018, ORESARA15, a PLATZ transcription factor, mediates leaf growth and senescence in *Arabidopsis*, In: *New Phyologist*, Vol. 220, Issue 2, p.609-623, DOI 10.1111/nph.15291

Kodym A., Zapata-Arias F.J., 1998, Natural light as an alternative light source for the *in vitro* culture of banana (*Musa acuminata* cv. 'Grande Naine'), In: *Plant Cell, Tissue and Organ Culture*, Vol. 55, Issue 2, p.141-145, DOI 10.1023/A:1006119114107

Krishna H., Alizadeh M., Singh D., Singh U., Chauhan N., Eftekhari M., Sadh R.K., 2016, Somaclonal variations and their applications in horticultural crops improvement, In: *3 Biotech*, Vol. 6, Issue 1, Article 54, 18 p., DOI 10.1007/s13205-016-0389-7

Kurz A., Lampel S., Nickolenko J.E., Bradl J., Benner A., Zirbel R.M., Cremer T., Lichter P., 1996, Active and inactive genes localize preferentially in the periphery of chromosome territories, In: *Journal of Cell Biology*, Vol. 135, n°5, p.1195-1205, DOI 10.1083/jcb.135.5.1195

Kusnetsov V., Landsberger M., Meurer J., Oelmüller R., 1999, The Assembly of the CAAT-box Binding Complex at a Photosynthesis Gene Promoter Is Regulated by Light, Cytokinin, and the Stage of the Plastids, In: *The Journal of Biological Chemistry*, Vol. 274, n°50, p.36009-36014, DOI 10.1074/jbc.274.50.36009

Lee-Huang S., Linb J.-J., Kung H.-F, Lin Huang P., Lee L., Lee Huang P., 1993, The human erythropoietin-encoding gene contains a CAAT box, TATA boxes and other transcriptional regulatory elements in its 5' flanking region, In: *Gene*, Vol. 128, Issue 2, p.227-236, DOI 10.1016/0378-1119(93)90567-M

Leva A.R., Petruccelli R., Rinaldi L.M.R., Somaclonal Variation in Tissue Culture: A Case Study with Olive, In: Leva A.R., Rinaldi L.M.R. (eds), Recent Advances in Plant in vitro Culture, 2012, p.123-150, DOI 10.5772/50367

Li C., Li Y., Bai L., Zhang T., He C., Yan Y., Yu X., 2013, Grafting-responsive miRNAs in cucumber and pumpkin seedlings identified by high-throughput sequencing at whole genome level, In: *Physiologia Plantarum*, Vol. 151, Issue 4, p. 406-422, DOI 10.1111/ppl.12122

Li J., Wu Z., Cui L., Zhang T., Guo Q., Xu J., Jia L., Lou Q., Huang S., Li Z., Chen J., 2014, Transcriptome Comparison of Global Distinctive Features Between Pollination and Parthenocarpic Fruit Set Reveals Transcriptional Phytohormone Cross-Talk in Cucumber (*Cucumis sativus* L.), In: *Plant and Cell Physiology*, Vol. 55, Issue 7, p.1325-1342, DOI 10.1093/pcp/pcu051

Lohmann C., Eggers-Schumacher G., Wunderlich M., Schöffl F., 2004, Two different heat shock transcription factors regulate immediate early expression of stress genes in *Arabidopsis*, In: *Molecular Genetics and Genomics*, Vol. 271, Issue 1, p.11-21, DOI 10.1007/s00438-003-0954-8

Long S.P., Humphries S., Falkowski P.G., 1994, Photoinhibition of Photosynthesis in Nature, In: *Annual Review of Plant Physiology and Plant Molecular Biology*, Vol. 45, p.633-662, DOI 10.1146/annurev.pp.45.060194.003221

Lorenzo O., Chico J. M., Sánchez-Serrano J. J., Solano R., 2004, *JASMONATE-INSENSITIVE1* Encodes a MYC Transcription Factor Essential to Discriminate between Different Jasmonate-Regulated Defense Responses in Arabidopsis, In: *The Plant Cell*, Vol. 16, Issue 7, p.1938-1950, DOI 10.1105/tpc.022319

Lu H.W., Miao H., Tian G.L., Wehner T.C., Gu X.F., Zhang S.P., 2015, Molecular mapping and candidate gene analysis for yellow fruit flesh in cucumber, In: *Molecular Breeding*, Vol. 35, n°64, 8 p., DOI 10.1007/s1103

Ładyżyński M., Burza W., Malepszy S., 2002, Relationship between somaclonal variation and type of culture in cucumber, In: *Euphytica*, Vol. 125, Issue 3, p.349-356, DOI 10.1023/A:101601782590

Malepszy S., Cucumber (*Cucumis sativus* L.), In: Bajaj Y.P.S. (ed), Biotechnology in Agriculture and Forestry 6. Crops II, 1988, p.277-293

Malepszy S., Nadolska-Orczyk A., 1989, *In vitro* Culture of *Cucumis sativus*. VIII. Variation in the Progeny of Phenotypically Not Altered R1 Plants, In: *Plant Breeding*, Vol. 102, Issue 1, p.66-72, DOI 10.1111/j.1439-0523.1989.tb00316.x

Malepszy S., Niemirowicz-Szczytt K., 1991, Sex determination in cucumber (*Cucumis sativus*) as a model system for molecular biology, In: *Plant Science*, Vol. 80, Issues 1-2, p. 39-47, DOI 10.1016/0168-9452(91)90271-9

Malepszy S., Burza W., Smiech M., 1996, Characterization of a cucumber (*Cucumis sativus* L.) somaclonal variant with paternal inheritance, In: *Journal of applied Genetics*, Vol. 37, n°1, p.65-78

Mao W., Li Z., Xia X., Li Y., Yu J., 2012, A Combined Approach of High-Throughput Sequencing and Degradome Analysis Reveals Tissue Specific Expression of MicroRNAs and Their Targets in Cucumber, In: *PLoS ONE*, Vol. 7, Issue 3, 10 p., DOI 10.1371/journal.pone.0033040

Martin C., Paz-Ares J., 1997, MYB transcription factors in plants, In: *Trends in Genetics*, Vol. 13, Issue 2, p.67-73, DOI 10.1016/S0168-9525(96)10049-4

Martín C., Uberhuaga E., Pérez C., 2002, Application of RAPD markers in the characterisation of Chrysanthemum varieties and the assessment of somaclonal variation, In: Euphytica, Vol. 127, Issue 2, p.247-253, DOI 10.1023/A:1020215016347

Martínez G., Forment J., Llave C., Pallás V., Gómez G., 2011, High-Throughput Sequencing, Characterization and Detection of New and Conserved Cucumber miRNAs, In: *PLoS ONE*, Vol. 6, Issue 5, 11 p., DOI 10.1371/journal.pone.0019523

Matheka J.M., Magiri E., Rasha A.O., Machuka J., 2008, In vitro Selection and Characterization of Drought Tolerant Somaclones of Tropical Maize (*Zea mays* L.), In: *Biotechnology*, Vol. 7, n°8, p.641-650, DOI 10.3923/biotech.2008.641.650

McWatters H.G., Bastow R.M., Hall A., Millar A.J., 2000, The *ELF3 zeitnehmer* regulates light signalling to the circadian clock, In: *Nature*, Vol. 408, p.716-720, DOI 10.1038/35047079

Menkens A.E., Schindler U., Cashmore A.R., 1995, The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins, In: *Trends in Biochemical Sciences*, Vol. 20, Issue 12, p.506-510, DOI 10.1016/S0968-0004(00)89118-5

Miao H., Zhang S., Wang X., Zhang Z., Li M., Mu S., Cheng Z., Zhang R., Huang S., Xie B., Fang Z., Zhang Z., Weng Y., Gu X., 2011, A linkage map of cultivated cucumber (*Cucumis sativus* L.) with 248 microsatellite marker loci and seven genes for horticulturally important traits, In: *Euphytica*, Vol. 182, Issue 2, p.167–176, DOI 10.1007/s10681-011-0410-5

Morcillo F., Gagneur C., Adam H., Richaud F., Singh R., Cheah S.-C., Rival A., Duval Y., Tregear J.W., 2006, Somaclonal variation in micropropagated oil palm. Characterization of two novel genes with enhanced expression in epigenetically abnormal cell lines and in response to auxin, In: *Tree Physiology*, Vol. 26, Issue 5, p.585-594, DOI 10.1093/treephys/26.5.585

Mróz T., Eksperymentalna walidacja danych RNA seq mutanta MSC19 ogórka (*Cucumis sativus* L.), Master's Thesis in Biotechnology: Szkoła Główna Gospodarstwa Wiejskiego w Warszawie, 2019, 65 p.

Mróz T.L., Ziółkowska A., Gawroński P., Pióro-Jabrucka E., Kacprzak S., Mazur M., Malepszy S., Bartoszewski G., 2015, Transgenic cucumber lines expressing the chimeric pGT::*Dhn24* gene do not show enhanced chilling tolerance in phytotron conditions, In: *Plant Breeding*, Vol. 134, Issue 4, p.468-476, DOI 10.1111/pbr.12275

Musharige T., Skoog F., 1962, A revised medium for rapid growth and bio assays with tobacco tissue cultures, In: *Physiologia Plantarum*, Vol. 15, Issue 3, p.473-497, DOI 10.1111/j.1399-3054.1962.tb08052.x

Noguero M., Atif R.M., Ochatt S., Thompson R.D., 2013, The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants, In: *Plant Science*, Vol. 209, p.32-45, DOI 10.1016/j.plantsci.2013.03.016

O'Rourke J.A., Yang S.S., Miller S.S., Bucciarelli B., Liu J., Rydeen A., Bozsoki Z., Uhde-Stone C., Jin Tu Z., Allan D., Gronwald J.W., Vance C.P., 2013, An RNA-Seq Transcriptome Analysis of Orthophosphate-Deficient White Lupin Reveals Novel Insights into Phosphorus Acclimation in Plants, In: *Plant Physiology*, Vol. 161, Issue 2, p.705-724, DOI 10.1104/pp.112.209254

Pawełkowicz M., Zieliński K., Zielińska D., Pląder W., Yagi K., Wojcieszek M., Siedlecka E., Bartoszewski G., Skarzyńska A., Przybecki Z., 2016, Next generation sequencing and omics in cucumber (*Cucumis sativus* L.) breeding directed research, In: *Plant Science*, Vol. 242, p.77-88, DOI 10.1016/j.plantsci.2015.07.025

Pawełkowicz M., Pryszcz L., Skarzyńska A., Wóycicki R.K., Posyniak K., Rymuszka J., Przybecki Z., Pląder W., 2019, Comparative transcriptome analysis reveals new molecular pathways for cucumber genes related to sex determination, In: *Plant Reproduction*, Vol. 32, Issue 2, p.193-216, DOI 10.1007/s00497-019-00362-z

Perez-de-Castro A.M., Vilanova S., Canizares J., Pascual L., Blanca J.M., Diez M.J., Prohens J., Pico B., 2012, Application of Genomic Tools in Plant Breeding, In: *Current Genomics*, Vol. 13, n°3, p.179-195, DOI 10.2174/138920212800543084

Phillips R.L., Kaeppler S.M., Olhoft P., 1994, Genetic instability of plant tissue cultures: breakdown of normal controls, In: *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 91, n°12, p.5222-5226, DOI 10.1073/pnas.91.12.5222

Piagnani C., Iacona C., Intrieri M.C., Muleo R., 2002, A New Somaclone of Prunus Avium Shows Diverse Growth Pattern under Different Spectral Quality of Radiation, In: *Biologia Plantarum*, Vol. 45, Issue 1, p.11-17, DOI 10.1023/A:1015182608782

Pląder W., Malepszy S., Burza W., Rusinowski Z., 1998, The relationship between the regeneration system and genetic variability in the cucumber (*Cucumis sativus* L.), In: *Euphytica*, Vol. 103, Issue 1, p.9-15, DOI 10.1023/A:1018359726626

Prinsi B., Negri A.S., Espen L., Piagnani M.C., 2016, Proteomic Comparison of Fruit Ripening between 'Hedelfinger' Sweet Cherry (*Prunus avium* L.) and Its Somaclonal Variant 'HS', In: *Journal of Agricultural and Food Chemistry*, Vol. 64, n°20, p.4171-4181, DOI 10.1021/acs.jafc.6b01039

Puranik S., Sahu P.P., Srivastava P.S., Prasad M., 2012, NAC proteins: regulation and role in stress tolerance, In: *Trends in Plant Science*, Vol. 17, Issue 6, p.369-381, DOI 10.1016/j.tplants.2012.02.004

Ramakers C., Ruijter J.M., Lekanne Deprez R.H., Moorman A.F.M., 2003, Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data, In: *Neuroscience Letters*, Vol. 339, Issue 1, p.62-66, DOI 10.1016/S0304-3940(02)01423-4

Reyes J.C., Muro-Pastor M.I., Florencio F.J., 2004, The GATA Family of Transcription Factors in Arabidopsis and Rice, In: *Plant Physiology*, Vol. 134, Issue 4, p.1718-1732, DOI 10.1104/pp.103.037788

Riechmann J.L., Ratcliffe O.J., 2000, A genomic perspective on plant transcription factors, In: *Current Opinion in Plant Biology*, Vol. 3, Issue 5, p.423-435, DOI 10.1016/S1369-5266(00)00107-2

del Río L.A., 2015, ROS and RNS in plant physiology: an overview, In: *Journal of Experimental Botany*, Vol. 66, Issue 10, p.2827-2837, DOI 10.1093/jxb/erv099

Rombauts S., Déhais P., Van Montagu M., Rouzé P., 1999, PlantCARE, a plant cis-acting regulatory element database, In: *Nucleic Acids Research*, Vol. 27, Issue 1, p.295-296, DOI 10.1093/nar/27.1.295

Shah S.N.M., Gong Z.-H., Arisha M.H., Khan A., Tian S.-L., 2015, Effect of ethyl methyl sulfonate concentration and different treatment conditions on germination and seedling growth of the cucumber cultivar Chinese long (9930), In: *Genetics and Molecular Research*, Vol. 14, n°1, p.2440-2449, DOI 10.4238/2015.March.30.2

Shang Y., Ma Y., Zhou Y., Zhang H., Duan L., Chen H., Zeng J., Zhou Q., Wang S., Gu W., Liu M., Ren J., Gu X., Zhang S., Wang Y, Yasukawa K., Bouwmeester H.J., Qi X., Zhang Z., Lucas W.J., Huang S., 2014, Biosynthesis, regulation, and domestication of bitterness in cucumber, In: *Science*, Vol. 346, Issue 6213, p.1084-1088, DOI 10.1126/science.1259215

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T., 2003, Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks, In: *Genome Research*, Vol. 13, n°11, p.2498-2504, DOI 10.1101/gr.1239303

Sirhindi G., Sharma P., Arya P., Goel P., Kumar G., Acharya V., Singh A.K., 2016, Genome-wide characterization and expression profiling of TIFY gene family in pigeonpea (*Cajanus cajan* (L.) Millsp.) under copper stress, In: *Journal of Plant Biochemistry and Biotechnology*, Vol. 25, Issue 3, p.301-310, DOI 10.1007/s13562-015-0342-6

Skarzyńska A., Pawełkowicz M., Pląder W, Przybecki Z., The utility of optical detection system (qPCR) and bioinformatics methods in reference gene expression analysis, In: Romaniuk R.S. (ed), Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2016, Proceedings of SPIE, Vol. 10031, Article 30, 11 p., DOI 10.1117/12.2249147

Skarzyńska A., Kuśmirek W., Pawełkowicz M., Pląder W., Nowak R.M., Assembly of cucumber (*Cucumis sativus* L.) somaclones, In: Romaniuk R.S., Linczuk M. (eds), Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, Proceedings of SPIE, Vol. 10445, Article 34, 8 p., 2017, DOI 10.1117/12.2280630

Smulders M.J.M., de Klerk G.J., 2011, Epigenetics in plant tissue culture, In: *Plant Growth Regulation*, Vol. 63, Issue 2, p.137-146, DOI 10.1007/s10725-010-9531-4

Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K.P., Kuhn M., Bork P., Jensen L.J., von Mering C., 2015, STRING v10: protein-protein interaction networks, integrated over the tree of life, In: *Nucleic Acids Research*, Vol. 43, Database issue, p.D447-D452, DOI 10.1093/nar/gku1003

Szwacka M., Krzymowska M., Osuch A., Kowalczyk M.E., Malepszy S., 2002, Variable properties of transgenic cucumber plants containing the thaumatin II gene from

Thaumatococcus daniellii, In: *Acta Physiologiae Plantarum*, Vol. 24, Issue 2, p.173-185, DOI 10.1007/s11738-002-0009-5

Tholl D., Chen F., Petri J., Gershenzon J., Pichersky E., 2005, Two sesquiterpene synthases are responsible for the complex mixture of sesquiterpenes emitted from *Arabidopsis* flowers, In: *The Plant Journal*, Vol. 42, Issue 5, p.757-771, DOI 10.1111/j.1365-313X.2005.02417.x

Tóth R., Kevei É., Hall A., Millar A.J., Nagy F., Kozma-Bognár L., 2001, Circadian Clock-Regulated Expression of Phytochrome and Cryptochrome Genes in Arabidopsis, In: *Plant physiology*, Vol. 127, Issue 4, p.1607-1616, DOI 10.1104/pp.010467

Tuteja N., Mahajan S., 2007, Calcium Signaling Network in Plants, In: *Plant Signaling & Behavior*, Vol. 2, Issue 2, p.79-85, DOI 10.4161/psb.2.2.4176

Untergasser A., Nijveen H., Rao X., Bisseling T., Geurts R., Leunissen J.A.M., 2007, Primer3Plus, an enhanced web interface to Primer3, In: *Nucleic Acids Research*, Vol. 35, Issue suppl_2, p.W71-W74, DOI 10.1093/nar/gkm306

Voorrips R.E., 2002, MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs, In: *Journal of Heredity*, Vol. 93, Issue 1, p.77-78, DOI 10.1093/jhered/93.1.77

Wang L., Zhang B., Li J., Yang X., Ren Z., 2014, Ethyl Methanesulfonate (EMS)-Mediated Mutagenesis of Cucumber (*Cucumis sativus* L.), In: *Agricultural Sciences*, Vol. 5, p.716-721, DOI 10.4236/as.2014.58075

Wang S.-L., Ku S.S., Ye X.-G., He C.-F., Kwon S.Y., Choi P.S., 2015, Current status of genetic transformation technology developed in cucumber (Cucumis sativus L.), In: *Journal of Integrative Agriculture*, Vol. 14, Issue 3, p.469-482, DOI 10.1016/S2095-3119(14)60899-6

Wu T., Qin Z., Zhou X., Feng Z., Du Y., 2010, Transcriptome profile analysis of floral sex determination in cucumber, In: *Journal of Plant Physiology*, Vol. 167, Issue 11, p.905-913, DOI 10.1016/j.jplph.2010.02.004

Xu X., Xu R., Zhu B., Yu T., Qu W., Lu L., Xu Q., Qi X., Chen X., 2015, A high-density genetic map of cucumber derived from Specific Length Amplified Fragment sequencing (SLAF-seq), In: *Frontiers in Plant Science*, Vol. 5, Article 768, 8 p., DOI 10.3389/fpls.2014.00768

Xu Z.-S., Chen M., Li L.-C., Ma Y.-Z., 2008, Functions of the ERF transcription factor family in plants, In: *Botany*, Vol. 86, n°9, 969-977, DOI 10.1139/B08-041

Yang L., Koo D.-H., Li Y., Zhang X., Luan F., Havey M.J., Jiang JK., Weng Y., 2012, Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly, In: *The Plant Journal*, Vol. 71, Issue 6, p.895-906, DOI 10.1111/j.1365-313X.2012.05017.x

Yang L., Li D., Li Y., Gu X., Huang S., Garcia-Mas J., Weng Y., 2013, A 1,681-locus consensus genetic map of cultivated cucumber including 67 NB-LRR resistance gene homolog and ten gene loci, In: *BMC Plant Biology*, Vol. 13, Article 53, 14 p., DOI 10.1186/1471-2229-13-53

Ye X., Song T., Liu C., Feng H., Liu Z., 2015, Identification of fruit related microRNAs in cucumber (*Cucumis sativus* L.) using high-throughput sequencing technology, In: *Hereditas*, Vol. 151, Issue 6, p.220-228, DOI 10.1111/hrd2.00057

Yoshida T., Fujita Y., Sayama H., Kidokoro S., Maruyama K., Mizoi J., Shinozaki K., Yamaguchi-Shinozaki K., 2010, AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation, In: *The Plant Journal*, Vol. 61, Issue 4, p.672-685, DOI 10.1111/j.1365-313X.2009.04092.x

Yusnita Y., Widodo W., Sudarsono S, 2005, *In Vitro* Selection of Peanut Somatic Embryos on Medium Containing Culture Filtrate of *Sclerotium rolfsii* and Plantlet Regeneration, In: *HAYATI Journal of Biosciences*, Vol. 12, n°2, p.50-56

Zhang N., Zhang H.-J., Zhao B., Sun Q.-Q., Cao Y.-Y., Li R., Wu X.-X., Weeda S., Li L., Ren S., Reiter R.J., Guo Y.-D., 2014, The RNA-seq approach to discriminate gene expression profiles in response to melatonin on cucumber lateral root formation, In: *Journal of Pineal Research*, Vol. 56, Issue 1, p.39-50, DOI 10.1111/jpi.12095

Zhao W., Yang X., Yu H., Jiang W., Sun N., Liu X., Liu X., Zhang X, Wang Y., Gu X., 2015, RNA-Seq-Based Transcriptome Profiling of Early Nitrogen Deficiency Response in Cucumber Seedlings Provides New Insight into the Putative Nitrogen Regulatory Network, In: *Plant and Cell Physiology*, Vol. 56, Issue 3, p.455-467, DOI 10.1093/pcp/pcu172

Zhou Y., Hu L., Wu H., Jiang L., Liu S., 2017, Genome-Wide Identification and Transcriptional Expression Analysis of Cucumber Superoxide Dismutase (SOD) Family in Response to Various Abiotic Stresses, In: *Internationla Journal of Genomics*, Vol. 2017, Article ID 7243973, 14 p., DOI 10.1155/2017/7243973

ONLINE RESSOURCES

FAO, 2019, FAOSTAT. Crops [online], Available at: http://www.fao.org/faostat/en/#data/QC/visualize (Accessed 07/07/2019)

INFOAGRO SYSTEMS, SL, 2014, Cucumber growing (Part I). Practical guide for a professional and intensive production of cucumber, vegetable that belongs to the cucurbitaceous family [online], Available at: http://agriculture.infoagro.com/crops/cucumber-growing--part-i-/ (Accessed 07/07/2019)

Annex I: Thermocycler programs for cDNA synthesis, PCR and qPCR

(1) Program used for cDNA synthesis

TEMPERATURE	TIME
25°C	10 minutes
37°C	2 hours
85°C	5 minutes
4°C	∞

(2) PCR program

TEMPERATURE	TIME	STEP	CYCLE NUMBER
95°C	3 minutes	Initial denaturation	
95°C	20 seconds	Denaturation	34 cycles
58°C	30 seconds	Annealing	
72°C	30 seconds	Extending – cDNA synthesis	
72°C	5 minutes	Final extending	
4°C	∞		

(3) Gradient PCR program used to test the primers. In the thermocycler, during the annealing step, each column has a different temperature ranging from 49,8°C to 65,1°C.

TEMPERATURE	TIME	STEP	CYCLE NUMBER
95°C	3 minutes	Initial denaturation	
95°C	20 seconds	Denaturation	34 cycles
49,8°C < 51,1°C < 52,5°C < 54,3°C < 56,2°C < 58,3°C < 60,2°C < 62°C < 63,5°C < 65,1°C	30 seconds	Annealing	
72°C	30 seconds	Extending – DNA synthesis	
72°C	5 minutes	Final extending	
4°C	∞		

(4) qPCR program used to verify the RNA-seq data and do the melting curve

TEMPERATURE	TIME	STEP	CYCLE NUMBER
50°C	20 seconds	qPCR	
95°C	10 minutes		
95°C	15 seconds	qPCR	40 cycles
60°C	1 minute		
95°C	15 seconds	Melt Curve	
60°C	1 minute		
95°C	30 seconds		
60°C	15 seconds		

Annex II: List of chosen genes to verify the RNA-seq data by qPCR in S2 and S3 lines

N° \ Line	S2	S3
1	*Cucsat.PASA.G825*	*Cucsat.PASA.G2525*
2	*Cucsat.PASA.G1373*	*Cucsat.PASA.G3325*
3	*Cucsat.PASA.G2244*	*Cucsat.PASA.G5228*
4	*Cucsat.PASA.G6064*	*Cucsat.PASA.G6440*
5	*Cucsat.PASA.G6481*	*Cucsat.PASA.G8234*
6	*Cucsat.PASA.G8448*	*Cucsat.PASA.G10613*
7	*Cucsat.PASA.G17205*	*Cucsat.PASA.G20923*
8	*Cucsat.PASA.G17802*	*Cucsat.PASA.G149*
9	*Cucsat.PASA.G1539*	*Cucsat.PASA.G2308*
10	*Cucsat.PASA.G16815*	*Cucsat.PASA.G7422*
11	*Cucsat.PASA.G18285*	*Cucsat.PASA.G8944*
12	*Cucsat.PASA.G20995*	*Cucsat.PASA.G9960*
13	*Cucsat.PASA.G21221*	*Cucsat.PASA.G11136*
14	*Cucsat.PASA.G2239*	*Cucsat.PASA.G13532*
15	*Cucsat.PASA.G3723*	*Cucsat.PASA.G13651*
16	*Cucsat.PASA.G7236*	*Cucsat.PASA.G14249*
17		*Cucsat.PASA.G20744*

	Diplôme : Diplôme d'Ingénieur de l'Institut des Sciences Agronomiques, Agroalimentaires, Horticoles et du Paysage Spécialité : Horticulture Spécialisation / option : Science et ingénierie du végétal (SIV)/ Semences et plants : recherche et développement, production, commercialisation (SEPRO) Enseignant référent : Agnès GRAPIN
Auteur(s) : Estelle BYSTRZYCKI Date de naissance* : 26 septembre 1996 Nb pages : 45 Annexe(s) : 2 Année de soutenance : 2019	Organisme d'accueil : Szkoła Główna Gospodarstwa Wiejskiego w Warszawie Adresse : Nowoursynowska 166 02-787 WARSZAWA Maître de stage : Magdalena PAWEŁKOWICZ

Titre français : Caractérisation de deux lignées somaclonales de concombre (*Cucumis sativus* L.) par approche transcriptomique

Titre anglais : Characterization of two somaclonal lines of cucumber (*Cucumis sativus* L.) through a transcriptomic approach

Résumé : Le concombre a une haute valeur économique dans le monde et particulièrement dans les pays de l'est en Europe. Il a déjà été longuement étudié pour permettre aux sélectionneurs d'obtenir de nouvelles variétés. Les NGS ont d'ailleurs grandement facilité l'amélioration des plantes. Différentes sources de nouveau matériel végétal sont disponibles. Parmi elles, on peut trouver la variation somaclonale (mutations issues de la culture *in vitro*) qui est la cible de cette étude. Deux lignées somaclonales de concombre, S2 et S3, ont été étudiées. Les deux phénotypes ont été comparés. Une analyse RNA-seq a été réalisée pour déterminer les DEGs entre les lignées somaclonales et le type sauvage dont elles sont issues, B10. Quelques-uns des DEGs ont été vérifiés par qPCR avec un succès d'en moyenne 97%. Un réseau de protéines a été construit avec les DEGs ainsi que des cartes chromosomiques et des heatmaps. Une analyse de promoteur a été réalisée. Les conclusions des analyses portent à croire que le retard de croissance observé est dû à des gènes en lien avec le cycle cellulaire, les hormones et les facteurs de transcription, sous le contrôle de divers éléments promoteurs impliqués dans des processus développementaux. Le retard de floraison peut s'expliquer par des gènes en lien avec les hormones et des facteurs de transcription sous l'influence d'hormones, de la lumière et du cycle circadien. Cependant, rien n'a pour l'instant été identifié pour expliquer la différence de couleur de S3 ou la supposée différence entre les fruits de S2 et ceux de B10.

Abstract: Cucumber has a high economic value in the world and particularly in eastern countries in Europe. It was already widely studied, giving tools for the breeders to create new varieties. Incidentally, NGS made plant breeding way easier. Different sources of nex plant material are available. Among them, somaclonal variation (mutations caused by *in vitro* culture) can be found. They are the target of this study. Two cucumber somaclonal lines, S2 and S3, were studied. Both phenotypes were compared. An RNA-seq analysis was carried out to determine DEG between somaclonal lines and the wild type from which they derive, B10. Some DEGs were checked by qPCR with an average success of 97%. A protein network was built with the DEGs as well as chromosome maps and heatmaps. A promoter analysis was carried out. The conclusions of the analyses suggest that the observed delay of growth comes from genes related to cell cycle, hormones and transcription factors, under the control of diverse promoter elements implied in developmental processes. The delay of flowering can be explained by genes related to hormones and transcription factors under the influence of hormones, light and the circadian cycle. However, nothing was identified to explain the difference of color of S3 or the supposed difference between S2 and B10 fruit.

Mots-clés : concombre, variation somaclonale, RNA-seq, DEGs, phénotype, NGS, réseau de protéines, bioinformatique

Key Words: cucumber, somaclonal variation, RNA-seq, DEGs, phenotype, NGS, protein network, bioinformatics

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*