Cucumber (*Cucumis sativus* L.) genome sequencing & comparative analysis

Wóycicki R et al. 2011, The genome sequence of a northern European cucumber (*Cucumis sativus* L.) cultivar unravels evolutionary adaptation mechanisms in plants. PLoS ONE 6(7): e22728. doi:10.1371/journal.pone.0022728

Dept. of Genetics, Breeding & Biotechnology, Faculty of Horticulture & Landscape Architecture, Warsaw University of Life Sciences – SGGW, Warsaw, Poland

Presentation scheme

I. Introduction

II. Material and methods

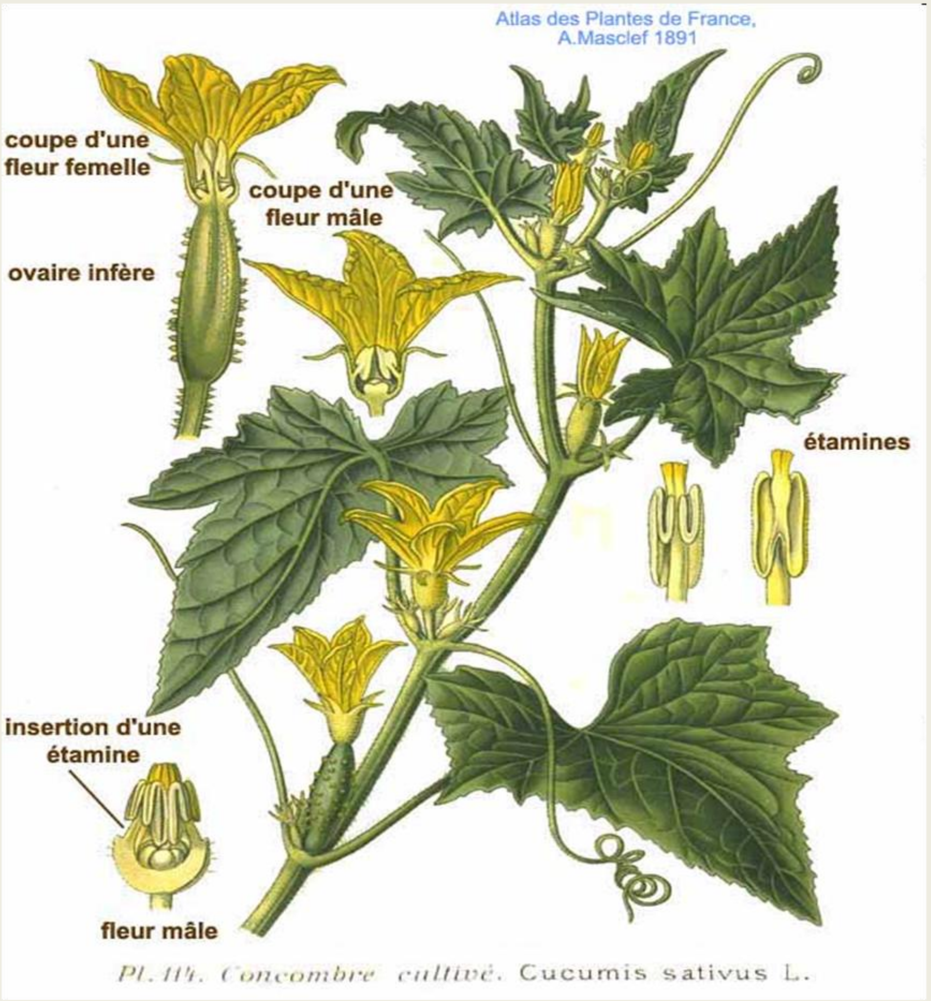
III. Results

- **1. Sequencing**
- **2. Genome reconstruction**
  - a. Reads assembly into contigs & scaffolds
  - b. Mapping of genome sequences onto chromosomes
- **3. Genome analysis**
  - a. Analysis of SSRs & other repeated sequences
  - b. Genome structural & functional annotation
  - c. Comparative analysis of genomes of two lines (B10 & 9930)
    - sequence level simillarity
    - differences in number of functional groups of genes
    - chromosomal rearrangements
  - d. Comparative analysis of gene promoters containing ABREs, DREs and EREs (CREs) between species (*A. thaliana*, *P. trichocarpa*, *O. sativa and C. sativus* lines B10 and 9930)
    - distribution & relative content of CREs
    - functional classification & analysis of genes containing CREs in their promoters.

V. Summary & Conclusions

Basic informations about cucumber

- Family: *Cucurbitaceae*
  Genus: *Cucurbita* i *Cucumis*
  Species: *Cucumis melo*,
  Cucumis sativus
- Economicallly importance
- Origin from Himalayas’ bottom
- Annual, outcrossing & monocieous
- 7 chromosomes pairs, diploid (2n)
- 367 000 000 bp (haploid genome)
- Model plant for basic and applied research


Atlas des Plantes de France,
A.Masclef 1891
coupe d'une fleur femelle
coupe d'une fleur mâle
ovaire infère
étamines
insertion d'une étamine
fleur mâle
Pl. 114. Concombre cultivé. Cucumis sativus L.

Materials & Methods

1. **Plant material –** young leaves of cucumber line B10
2. **Genomic DNA isolation -** GenElute Plant Genomic DNA Miniprep Kit (Sigma Aldrich, Buchs, Switzerland)
3. **65,260 BAC clones from two libraries** (HindIII (Gutman et al. 2008) & BamHI/MboI (Amplicon Express, Pullman, WA, USA))
4. **Sanger sequencing of 89’088 BESs** – (Agencourt Bioscience Corporation, Beverly, MA ,USA (2008) now Beckman Coulter Genomics)
5. **Bioinformatics analysis of BESs –** quality– Lucy, BLAST; SSRs- Phobos; other repeats- RepeatModeler, Repbase Update, BLAST, TIGR Plant Repeat Database
6. **454 Titanium pyrosequencing** – 12x genome coverage in single (8x) and PE (4x, 3000 pb) – Agencourt Bioscience Corporation

   reads quality– sffinfo , sff_extract

Materials & Methods

7. **Genome assembly using 454 and BES reads**

   - A version - Celera
   - B version - Celera oraz Arachne

8. **Quality check of assembled genome sequences**

   Simillarity of assembled genome sequences – MUMmer, RepeatMasker

   No. & homology of BESs, 63,035 EST unigenes, BAC & Fosmid clones' seqeunces to asambled contigs- BLAT, MUMmer, RepeatMasker, coverage of assembled contigs in reads after 454 Titanium

9. **Mapping of B10 & 9930 genomes onto chromosomes**

   1,883 molecular markers – BLAST, Arachne, MUMmer

10. **Structural annotation of genomes of B10 & 9930 lines**

    Gene prediction using the model made with GeneMark.hmm ES (Mark Borodovsky)

    Gene model & prediction veryfication using sequenices of 63,035 EST unigenes & 422 cDNAs – BLAT

Materials & Methods

11. Functional annotation of genomes of B10 & 9930 lines

Predicted peptides vs. GenBank db – BraGOMap (Wóycicki et al., 2008)

Gene Ontology classification –iProClass db, GORetriver

12. Comparative analysis of genome sequences of B10 and 9930 lines

Analysis of sequences simillarity together with SNPs/INDELs discovery - MUMmer

Comparison of sequences mapping onto chromosomes - Mauve, MUMmer

13. Comparative analysis of gene promotores between species (*A. thaliana*, *P. trichocarpa*, *O. sativa* and *C. sativus* lines B10 and 9930)

Identification of genes containing ABRE, DRE and ERE elements in their promoters (1,000 bp upstream the start codon (ATG)) – Patmatch

Comparison of protein sequences – BLAST, OrthoMCL

Functional classification of genes containing ABRE, DRE and ERE elements in their promoters – GOSlim

Materials & Methods

Identification of putative transcription factors from C. sativus line B10 and 9930 – PFAM, ClustalW, MEGA 4 (Neighbour-Joining method)

ABA treatment and electrolyte leakage - seedlings were subjected to 200µM Abscisic acid (ABA) (Sigma Aldrich, St. Louis, MO, USA) for 3 days. The last ABA treatment was made 3 hours before freezing treatment. Electrolyte leakage experiments were performed as previously described (Jaglo-Ottosen KR et al. 1998) with modifications. At least five replicates for each data point.

13. Computer power

3 computing stations – 28 SSP, 88 GB RAM (Applied Omics (Warsaw, Poland) & Warsaw University of Life Sciences - SGGW)

14. Home-made Perl scripts (rafal_woycicki.users.sggw.pl/rw_scripts.html)

Results

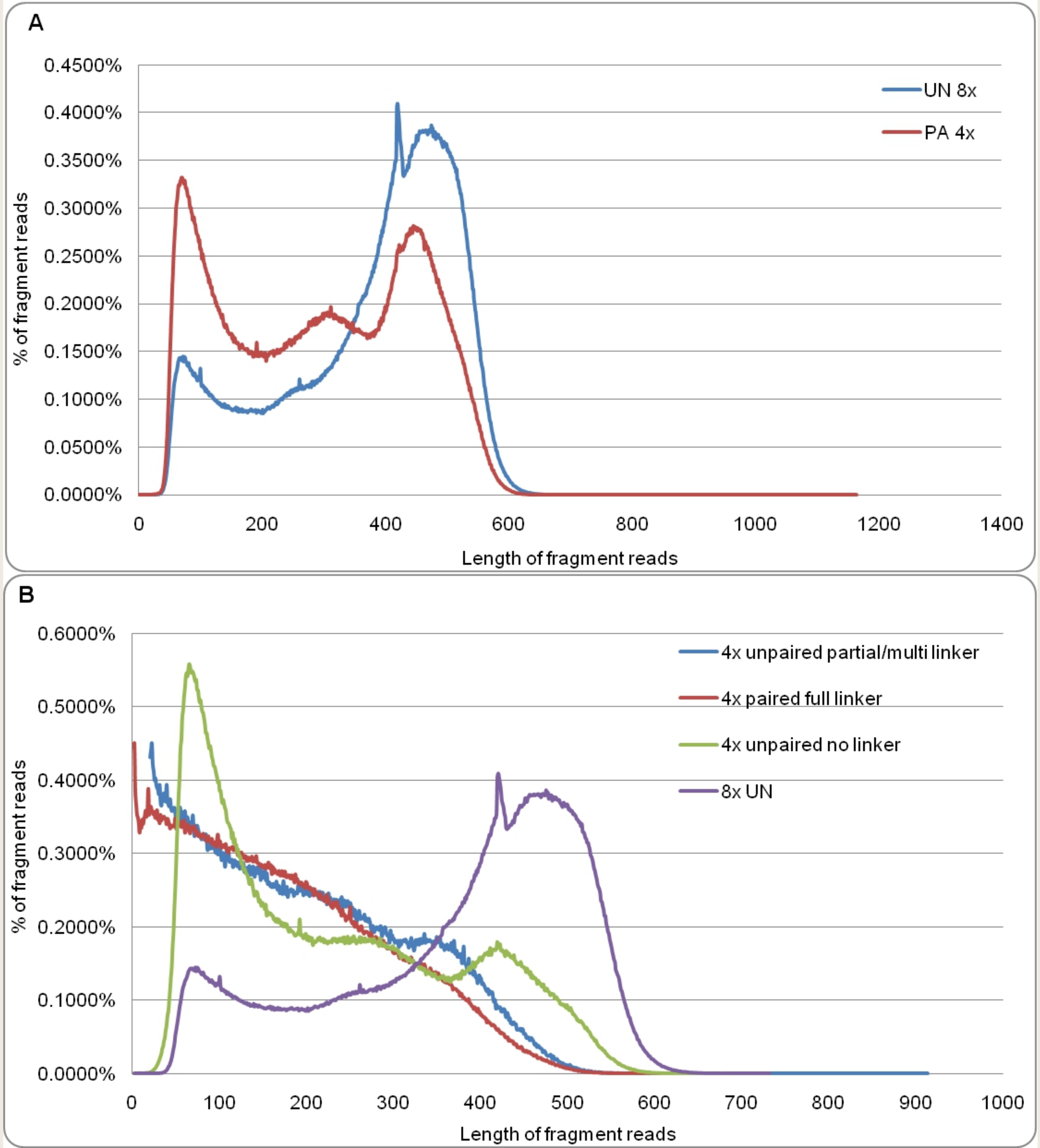
Sequencing

BESs sequencing

Feature	Sum	%
Totall no. of reads	**84,493**	**94.84**
Not-accepted reads	19,883	23.53
Goog quality BESs	**64,610**	**76.47**
Chloroplastom homology sequences	2,094	3.24
Mitochondrion homology sequences	297	0.46
Nuclear genome derived sequences	**62,220**	**71.98**
Mean lenght of nuclear BESs [nt]	**737**	
Sum lenght of nuclear BESs [nt]	45,563,499	
Genome in BESs [%]	**12.42**	

454 Titanium pyrosequencing

	4× paired full linker	4× unpaired no linker	**8× unpaired**	Summary
Total no. reads	**3,204,606**	3,999,255	**7,970,914**	**15,174,775**
< 100 nt	33.67%	26.86%	6.49%	ND
100-300 nt	49.55%	46.07%	20.14%	
301-500 nt	16.66%	25.47%	55.05%	
> 500 nt	0.12%	1.61%	18.32%	
Mean lenght	**171.53**	220.01	**374.00**	**290.66**
Lenght sum	549,690,047	879,890,390	2,981,159,897	4,410,740,334
Coverage	**1.50**	2.40	**8.12**	**12.02**


A
0.4500%
0.4000%
0.3500%
0.3000%
0.2500%
0.2000%
0.1500%
0.1000%
0.0500%
0.0000%
% of fragment reads
0
200
400
600
800
1000
1200
1400
Length of fragment reads
UN 8x
PA 4x
B
0.6000%
0.5000%
0.4000%
0.3000%
0.2000%
0.1000%
0.0000%
% of fragment reads
0
100
200
300
400
500
600
700
800
900
1000
Length of fragment reads
4x unpaired partial/multi linker
4x paired full linker
4x unpaired no linker
8x UN

Genome reconstruction

Genome reads assembly

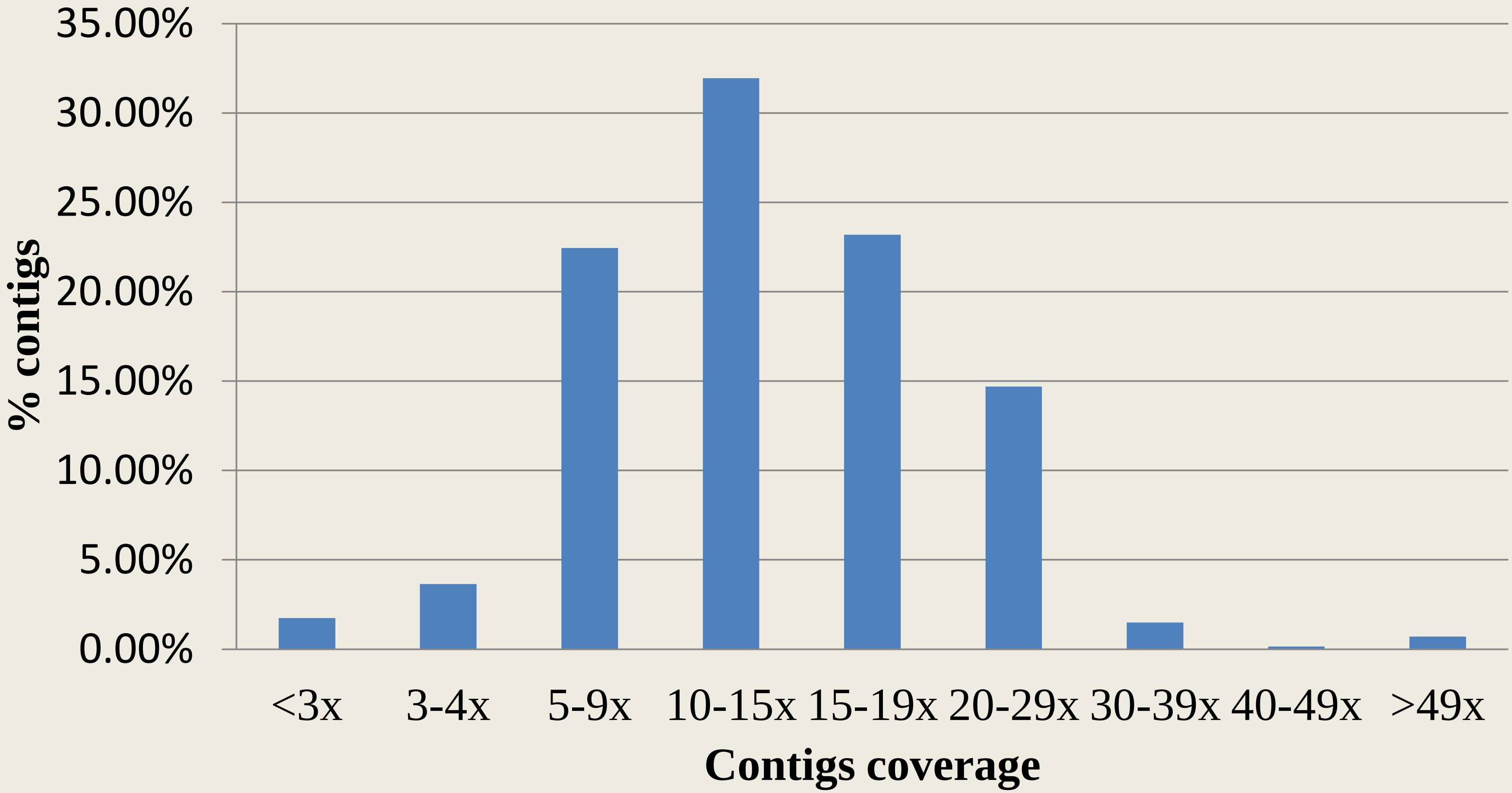
Cucumber genomes assembly results

Feature	**B10** **A version** 12x 454, BESs	**B10** **B version** 12x 454, BESs	**9930** (Huang et al. 2009) **72x Illumina/Sanger**	**GY14** (Miller et al. 2009, Cavagnaro et al 2010) **36x 454**
Contigs' length sum [Mbp]	**197**	**193**	**185**	**200**
% of genome size	53.79	52.64	50.61	54.50
No. contigs	**15,667**	**16,454**	**12,195**	**7,901**
Mean contigs length [bp]	**12,972**	**11,712**	**15,230**	**BD**
N50 contigs length [bp]	27,086	23,200	30,248	37,600
Scafolds length sum [bp]	**224**	**321**	**203**	**203**
% of genome size	61.24	87.82	55.39	55.33
No. scafolds	**4,173**	**13,116**	**1,792**	**3,610**
Mean scafolds lenght [bp]	54,070	24,500	113,435	48,129
N50 scafolds lenght [bp]	**2,324,038**	**315,056**	**1,509,230**	**993,000**

Genome assembly quality check

- More than 98% simillarity between both versions of genome assembly, about 10 Mbp differing those two versions
- 97% of 63,035 cucumber EST unigenes & other genome sequences (BACs, Fosmids) mapped with 98% simillarity
- 51,936 of BESs (83,48%) uniqly mapped to the contigs
- Almost 95% of assembled total contigs lenght are longer then mean gene length

Mean assembled genome coverega > 14x, 98% contigs with > 3x coverage



Mapping of assembled genome onto chromosomes

Genome onto chromosomes

Chromosome no.	**A version - Celera**		**B Version – Celera/Arachne**	
	Contigs sum length [Mbp]	Scafolds sum length [Mbp]	Contigs sum length [Mbp]	Scafolds sum length [Mbp]
1 (4)	20.87	28.60	20.30	34.91
2 (2)	20.81	25.16	20.24	34.86
3 (3)	34.89	39.09	33.96	60.83
4 (6)	26.81	33.16	26.13	47.75
5 (1)	23.18	29.98	22.59	45.04
6 (5)	24.05	30.31	23.50	45.98
7 (7)	16.47	20.44	15.97	31.14
Total	**167.11**	**206.73**	**162.73**	**300.53**
% assembled genome	**85.83**	**92.29**	**84.23**	**93.04**

Genome analysis

SSRs & other repeats analysis

SSRs characteristics

Nucleotides repeats motifs	Contigs[%]	BESs [%]
	0.95	**0.73**
Mono-	**9.83**	**25.91**
A	96.77	97.07
C	3.23	2.93
Di-	**24.97**	**20.15**
AT	72.12	69.63
AG	20.05	20.79
Tri--	**22.15**	**18.62**
AAT	47.17	42.23
AAG	31.68	32.51
Tetra-	**23.36**	**19.28**
AAAT	37.04	35.65
AAAG	19.25	20.46
Penta-	**9.63**	**7.7.**
AAAAG	26.44	30.12
AAAAT	22.36	17.71
Hexa-	**6.55**	**5.36**
AAAAAG	18.44	16.38
AAAAAT	8.74	6.51

Plant repeated elements analysis

Super-Class	**Class**	**Sub-Class**	**Contigs[%]**	**BESs [%]**
			17.82	**48.13**
Transposable Elements	Retrotransposons	Ty1-copia	1.78	3.25
		Ty3-gypsy	0.53	0.59
		LINE	0.46	1.02
		SINE	0.00	0.00
		Unclassified	2.80	3.79
	Transposons	Ac/Ds	0.01	0.01
		CACTA, En/Spm	0.20	0.19
		Mutator (MULE)	0.03	0.03
		Unclassified	0.55	0.82
	MITEs		0.00	0.00
Centromere related sequences			0.06	0.11
Telomere related sequences			0.00	0.02
rRNA genes	**45S rDNA**		**0.18**	**8.67**
	5S rDNA		**0.02**	**0.29**
Unclassified repeated sequences			1.00	0.90
Cucumber specific repeated sequences			10.09	23.81
Small RNAs			**0.11**	**4.63**

Genome annotation

Structural annotation

Gene prediction results

Feature	**Line B10**	**Line 9930**	**Line 9930 (Huang et al. 2009)**
No. of protein coding genes	**26,587**	24,678	26,682
Mean length of exons [bp]	**201**	207	238
Mean no. of exons per gene	**5.49**	5.79	4.39
Mean intron length [bp]	**436**	441	483
Mean lenght of integenic region [bp]	3,009	2,864	ND
Mean lenght of coding sequence [bp]	1,103	1,198	1,046
Mean lenght of transcribed region [bp]	3,058	3,309	2,685
Mean gene lenght [bp]	**4,563**	4,741	ND

Functional annotation

Functional annotation results

Feature	Line B10		Line 9930	
	No.	**%**	**No.**	**%**
No. proteins >= 100 aa	23,190	87.22	21,177	85.81
Similarities in GenBank	19,562	84.36	ND	ND
Annotated genes in GenBank	**16,944**	**73.07**	**16,443**	**77.65**
Gene products with GO	**12,643**	**54.52**	**12,363**	**58.38**
GO Biological Process (BP)	9,015	71.30	8,813	71.29
GO Molecular Function (MF)	11,391	90.10	11,116	89.91
GO Cellular Compartment (CC)	5,355	42.36	5,239	42.38

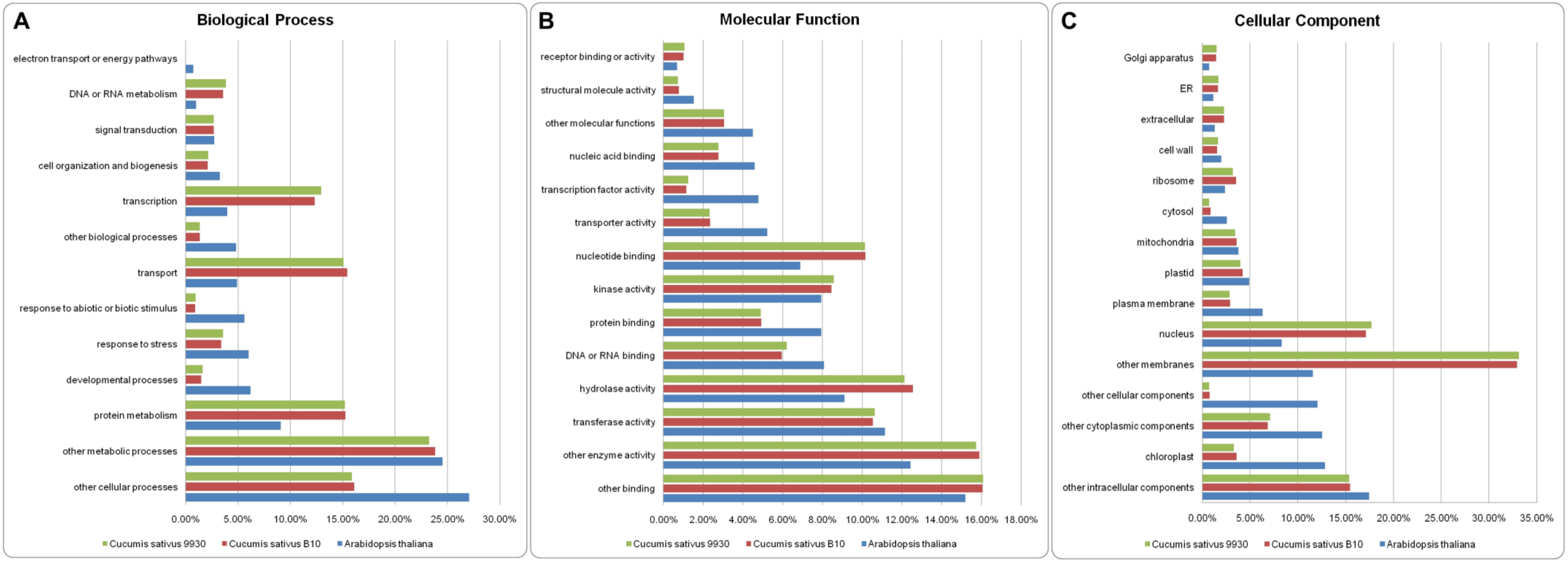
Comparative analysis of genomes of B10 and 9930 lines

Genomes sequence similarity, SNPs & INDELs

97,40% similarity of assembled parts of genomes of lines B10 and 9930

Feature	**No. SNP**	**SNP frequency for 1 Kbp**	**No. INDEL**	**INDEL frequency for 1 Kbp**
Whole assembled genome	**811,274**	**4.22**	**485,048**	**2.53**
Transcribed region	196,845	2.48	108,359	1.37
Exons	**45,449**	**1.60**	**16,584**	**0.58**
Gene promotor regions (-1000 from ATG)	79,716	3.85	49,856	2.40

Differences in functional group of genes

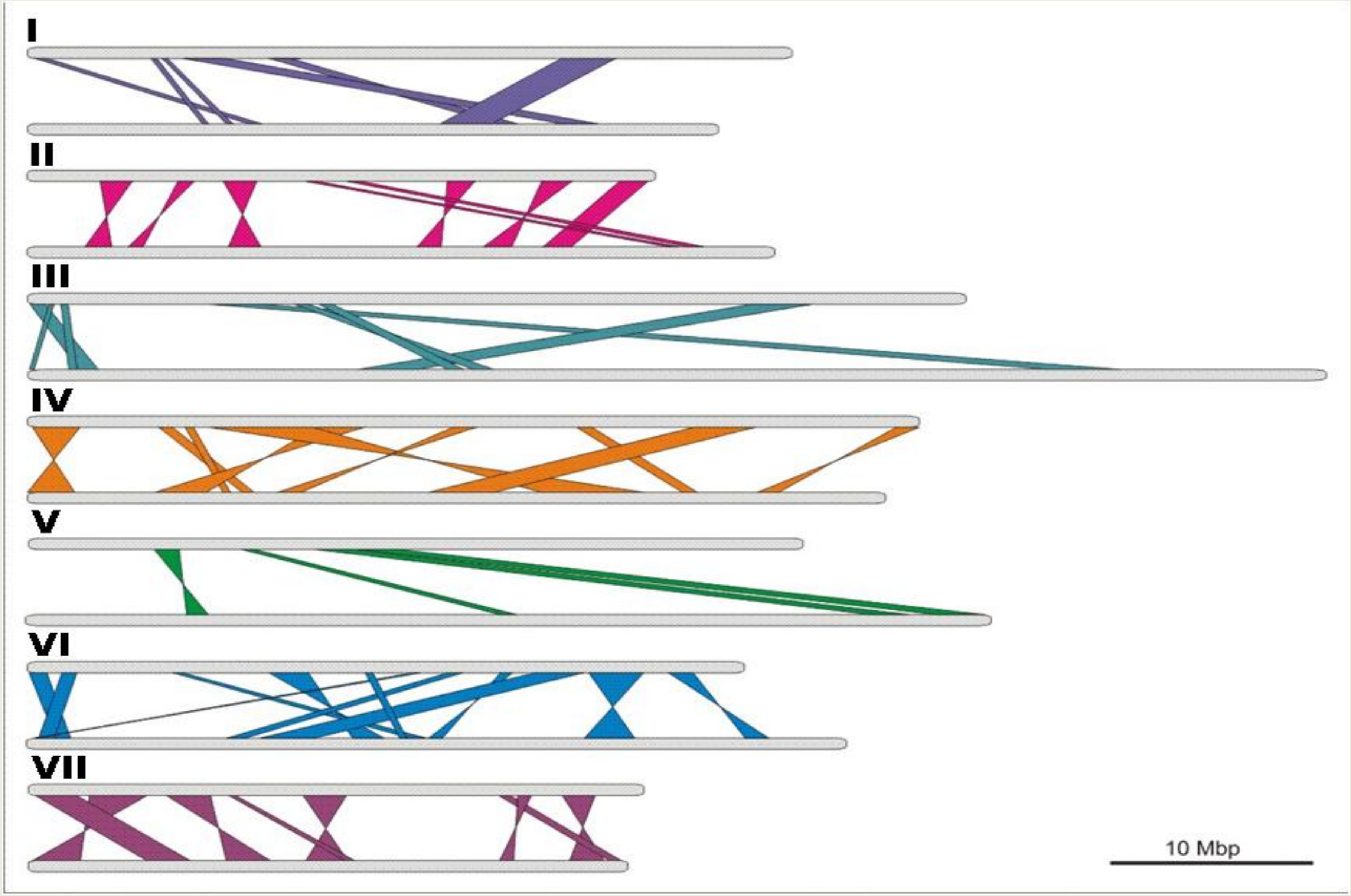

A
Biological Process
electron transport or energy pathways
DNA or RNA metabolism
signal transduction
cell organization and biogenesis
transcription
other biological processes
transport
response to abiotic or biotic stimulus
response to stress
developmental processes
protein metabolism
other metabolic processes
other cellular processes
0.00%
5.00%
10.00%
15.00%
20.00%
25.00%
30.00%
Cucumis sativus 9930
Cucumis sativus B10
Arabidopsis thaliana
B
Molecular Function
receptor binding or activity
structural molecule activity
other molecular functions
nucleic acid binding
transcription factor activity
transporter activity
nucleotide binding
kinase activity
protein binding
DNA or RNA binding
hydrolase activity
transferase activity
other enzyme activity
other binding
0.00%
2.00%
4.00%
6.00%
8.00%
10.00%
12.00%
14.00%
16.00%
18.00%
Cucumis sativus 9930
Cucumis sativus B10
Arabidopsis thaliana
C
Cellular Component
Golgi apparatus
ER
extracellular
cell wall
ribosome
cytosol
mitochondria
plastid
plasma membrane
nucleus
other membranes
other cellular components
other cytoplasmic components
chloroplast
other intracellular components
0.00%
5.00%
10.00%
15.00%
20.00%
25.00%
30.00%
35.00%
Cucumis sativus 9930
Cucumis sativus B10
Arabidopsis thaliana

Comparison of functional annotation results

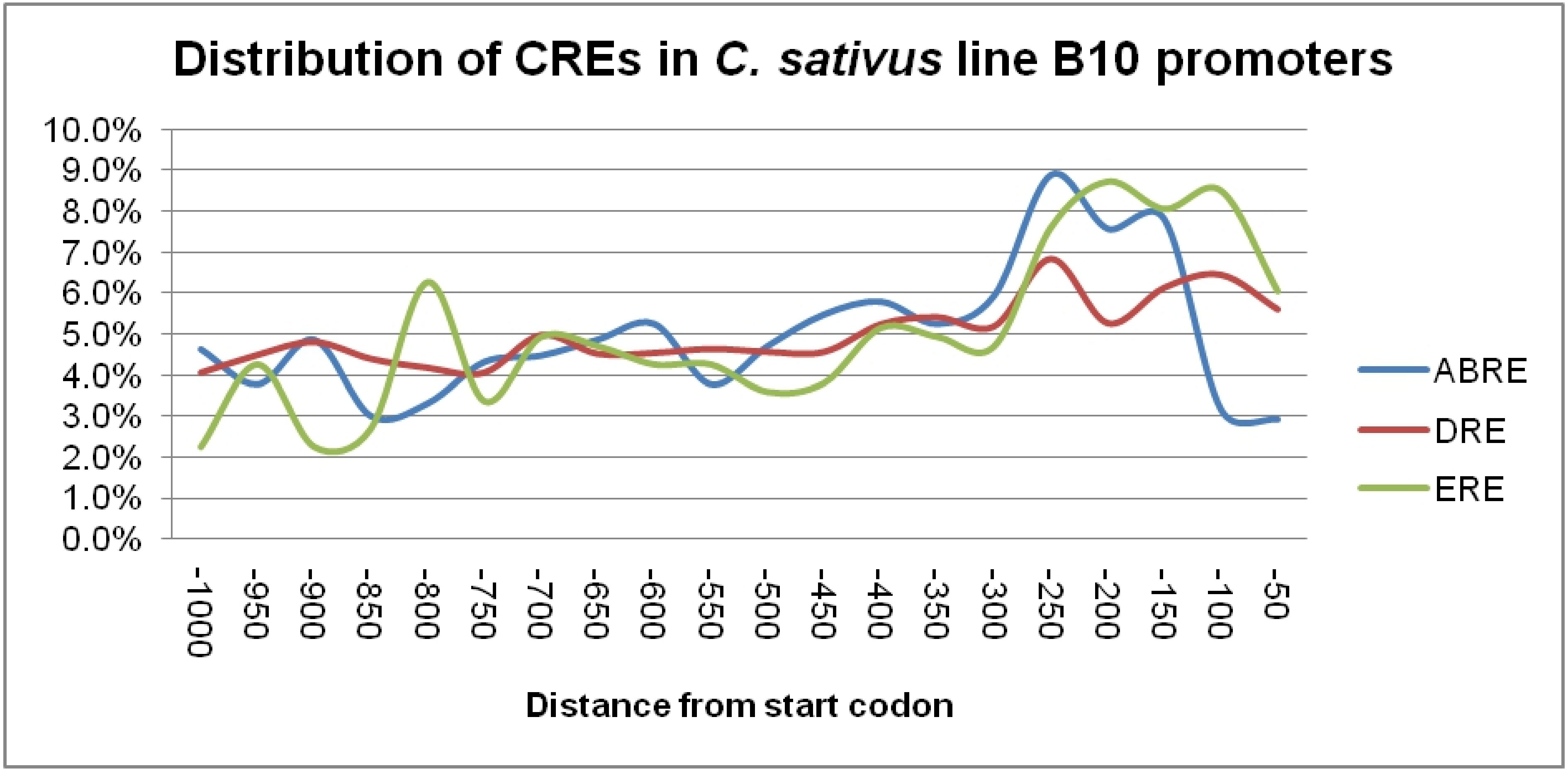
Process	**Line B10**	**Line 9930**	**Envinormental conditions**
Photosynthesis	+	-	Temper climate of Northern Europe: cold, low light intensity
Sugar metabolism	+	-	
Respiration	+	-	
Reg.of gene expreession	+	-	
Chlorophyll degradation	+	-	
Nitrogen binding as amonium ions	+	-	Continouos and higher emision of CO2 in Europe than in South-Eastern Asia, when counting from beginning of industrial era to 80's of XX century - lowered abillity for binding nitrogen ions
Oxidative stress resistance	-	+	Subtropical climate of South-Eastern China: high sesonal intensity of sun light including UV-B radiation, together with high temperature
High temperature resistance	-	+	

Chromosomal rearrangements

Chromosomal rearrangements visualization



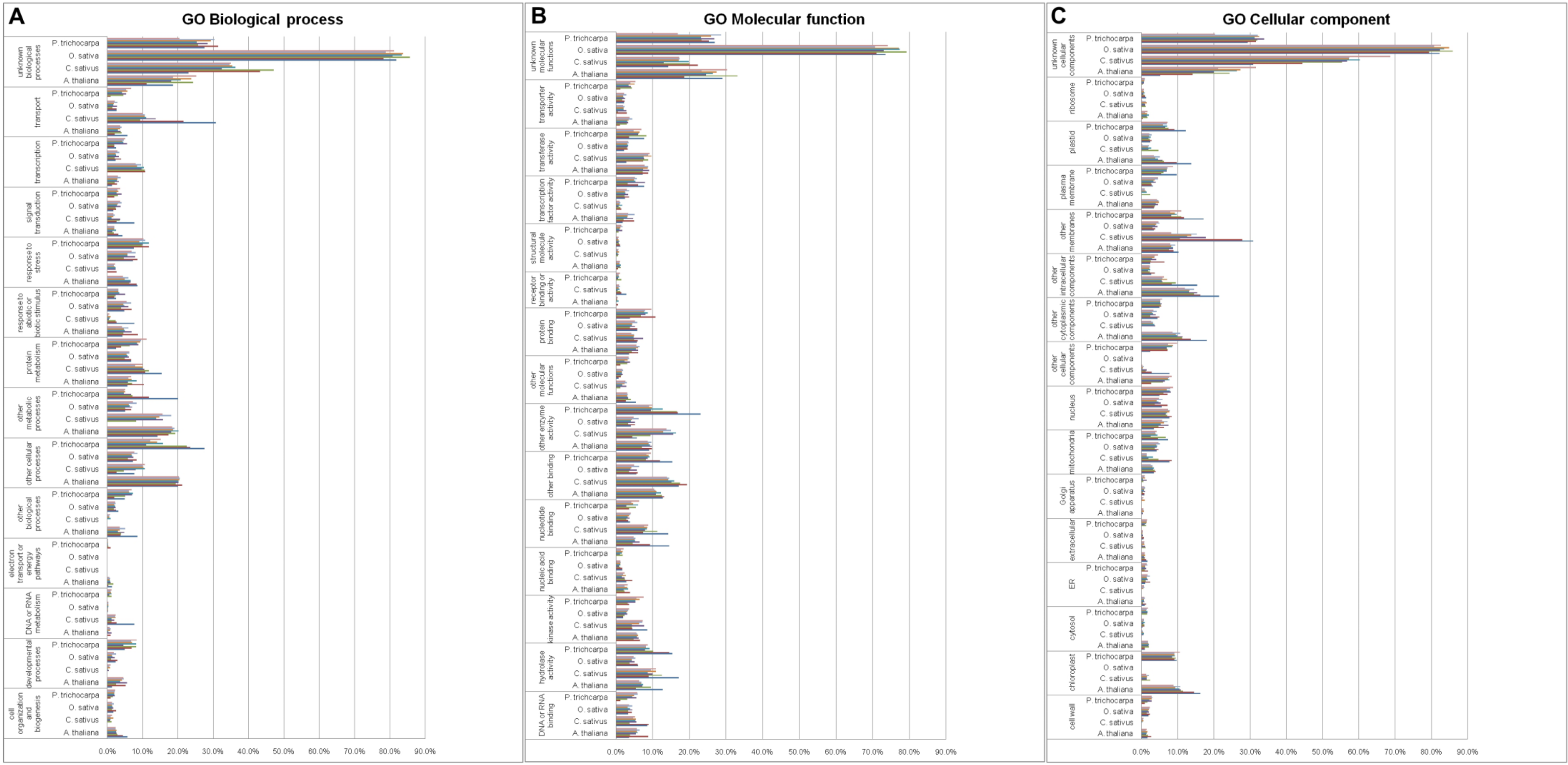
Comparative analysis of gene promoters containing ABREs, DREs and EREs between 4 species: *A. thaliana*, *P. trichocarpa*, *O. sativa* and *C. sativus* lines B10 and 9930


Distribution of CREs in *C. sativus* line B10 promoters
10.0%
9.0%
8.0%
7.0%
6.0%
5.0%
4.0%
3.0%
2.0%
1.0%
0.0%
-1000
-950
-900
-850
-800
-750
-700
-650
-600
-550
-500
-450
-400
-350
-300
-250
-200
-150
-100
-50
Distance from start codon
ABRE
DRE
ERE

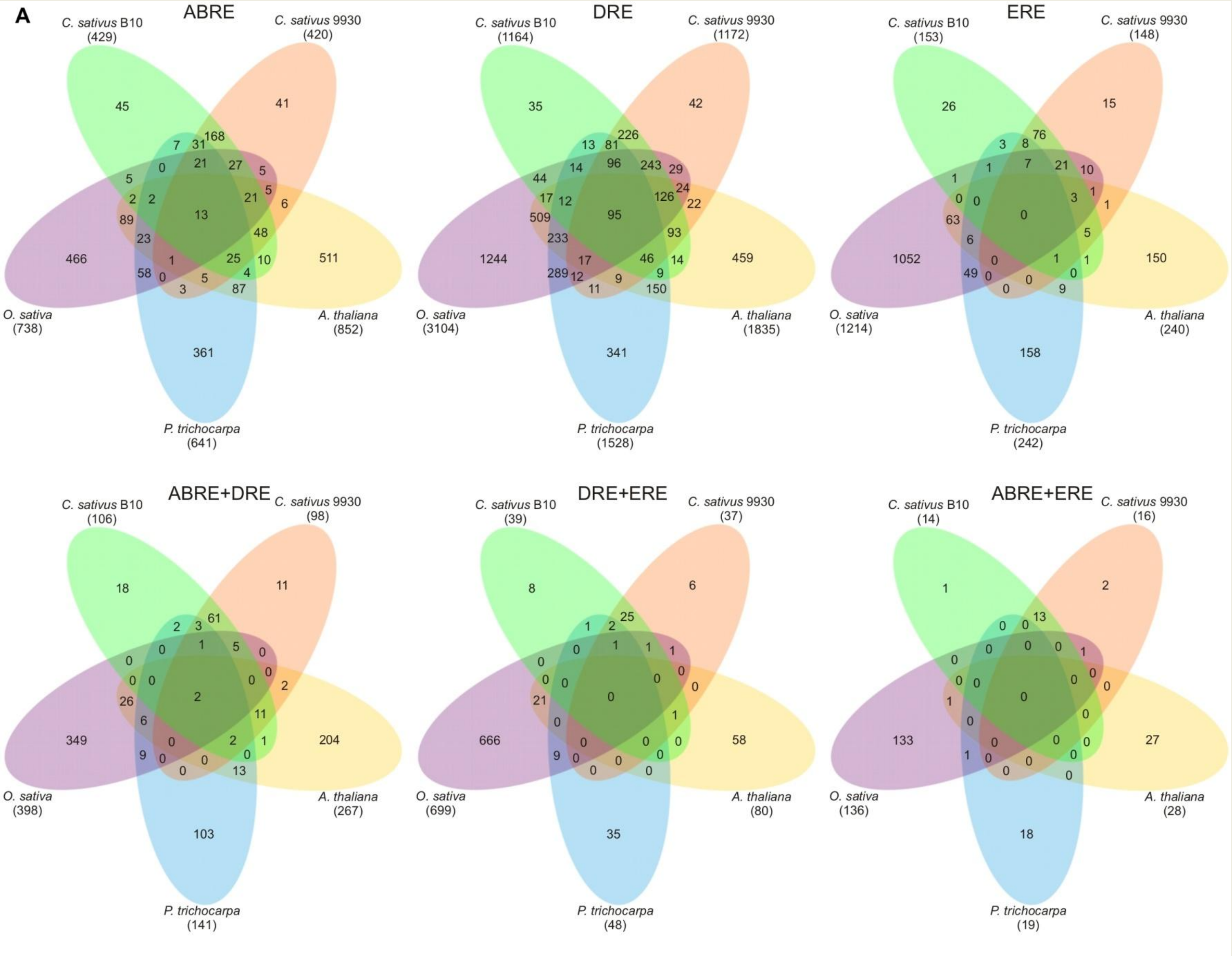
Relative content of ABREs, DREs and EREs in promoters of genes of 4 species

Species	ABRE			DRE			ERE		
	%	p-value	Av.	%	p-value	Av.	%	p-value	Av.
A. thaliana	24.8	2.86 e-05	1.1763	67.7	1.33 e-04	1.2098	7.4	7.42 e-06	1.037
C. sativus line B10	22.1	2.47 e-05	1.1174	70.3	1.08 e-04	1.1766	7.7	5.42 e-06	1.0251
C. sativus line 9930	22.4	2.47 e-05	1.1385	69.4	1.09 e-04	1.1885	8.2	5.46 e-06	1.0509
O. sativa	**8.0**	5.30 e-05	1.1121	73.7	3.45 e-04	1.6561	**18.3**	3.48 e-05	1.2503
P. trichocarpa	**19.1**	2.80 e-05	1.1234	70.8	1.27 e-04	1.3172	**10.1**	7.01 e-06	1.1154

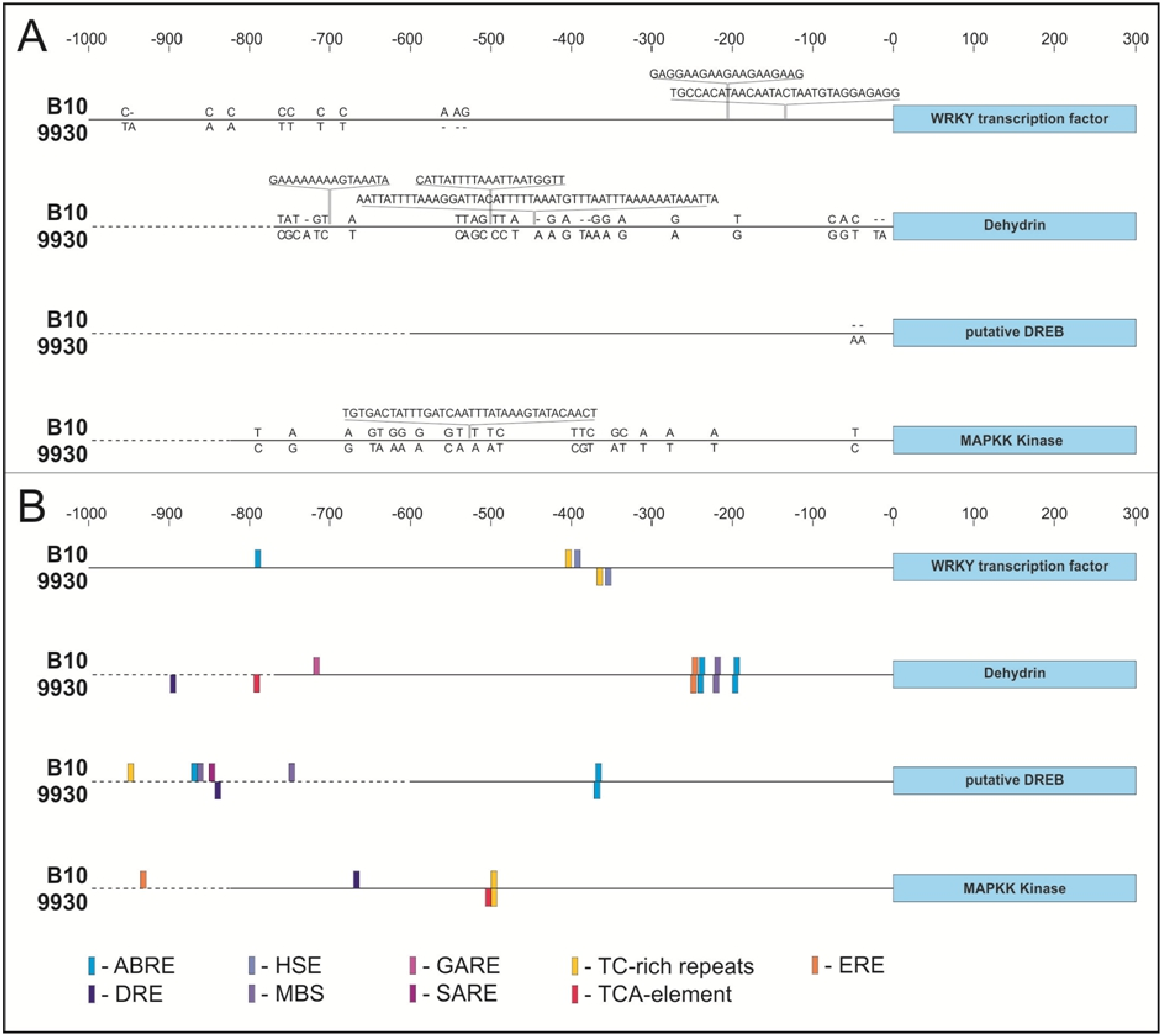
Functional classification of 4 species genes containing ABREs, DREs and EREs in their promoters.



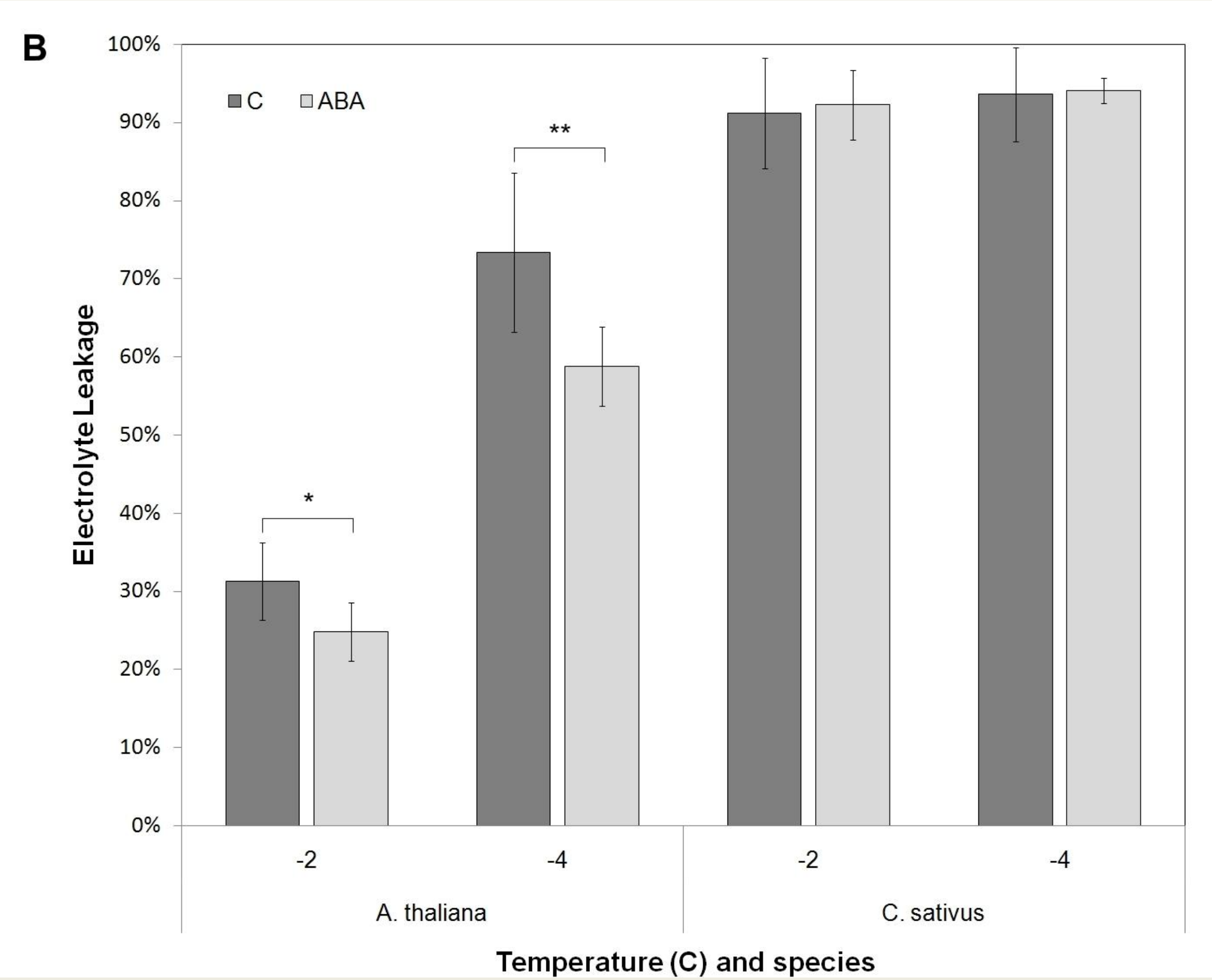
Functional analysis of genes containing ABREs, DREs and EREs in their promoters



Changes observed in promoter of selected orthologous genes of two *C. sativus* lines


A
-1000 -900 -800 -700 -600 -500 -400 -300 -200 -100 -0 100 200 300
GAGGAAGAAGAAGAAGAAG
TGCCACATAACAATACTAATGTAGGAGAGG
B10
9930
C- C C CC C C A AG
TA A A TT T T - - -
WRKY transcription factor
GAAAAAAAAGTAAATA
CATTATTTTAAATTAATGGTT
AATTATTTTAAAGGATTACATTTTTAAATGTTTAATTTAAAAAATAAATTA
B10
9930
TAT - GT A TTAG TT A - G A - -GG A G T C A C - -
CGC A TC T CAGC CC T A A G TAA A G A G G G T TA
Dehydrin
B10
9930
- -
AA
putative DREB
TGTGACTATTTGATCAATTTATAAAGTATACAACT
B10
9930
T A A GT GG G G T T T C TTC GC A A A T
C G G TA AA A C A A AT CGT AT T T T C
MAPKK Kinase
B
-1000 -900 -800 -700 -600 -500 -400 -300 -200 -100 -0 100 200 300
B10
9930
WRKY transcription factor
B10
9930
Dehydrin
B10
9930
putative DREB
B10
9930
MAPKK Kinase
- ABRE
- DRE
- HSE
- MBS
- GARE
- SARE
- TC-rich repeats
- TCA-element
- ERE

Freezing tolerance tests of non-acclimated *A. thaliana* and *C. sativus* seedlings after ABA treatment



Summary

- ✓ **62’220 Sanger sequenced nuclear cucumber BESs - 12,49% genome size**
- ✓ **454 Titanium *de-novo*** sequencing of cucumber genome (*Cucumis sativus* L.) line B10 with **12x read coverage**
- ✓ **52% of genome assembled** into contigs**,** 48% of the cucumber genome consists of plant repeats (basing on BESs)
- ✓ **26,587 of gene structures** were predicted *de-novo*
- ✓ **12,643 proteins (47,55%)** with Gene Ontology
- ✓ **85% of assembled contigs & 93% of scaffolds** were mapped onto chromosomes

Summary

- ✓ **Differences in Gene Copy Numbers**, explaining envinormental adaptations, were reported between two cucumber lines originating from two diverse envinorments

- ✓ **Global Intra-Chromosomal Rearrangements** of inversions & translocations of large regions, which could lead to envinormental adaptations of two cucumber lines, were reported

- ✓ **Substantial differences in CRE content** between all analyzed species and varieties (*C. sativus* (B10 and 9930)).

- ✓ **Only a small fraction of the groups of orthologous genes** with the highest sequence similarity in analyzed 4 species **have the same CRE profiles in their promoters.**

- ✓ ABA-treatment experiments together with *in silico* analysis of CRE shuffling explains why *C. sativus* is much more susceptible for cold and chilling stresses than *A. thaliana*.

Conclusion

Eukaryotic organisms are equipped with a high degree of freedom with respect to:

1. Variability of promoters in terms of regulatory elements,

2. (Intra-) Chromosomal rearrangements ,

that allow for formation of new lines/varieties and species adapted to new ecological niches.

Grants:

- Polish Ministry of Science & Higher Education: PBZ-MNiSW-2/3/2006/36, NN302429734, NN302363333,
- Foundation for Polish Science Welcome Program / European Regional Development Fund: Welcome 2008/1
- U.S. NIH: GM47853

Warsaw University of Life Sciences - - SGGW

Georgia Tech

People:

Justyna Witkowicz
Piotr Gawroński
Joanna Dąbrowska
Alexandre Lomsadze
Magdalena Pawełkowicz
Ewa Siedlecka
Kohei Yagi
Wojciech Pląder
Anna Seroczyńska
Mieczysław Śmiech
Wojciech Gutman
Katarzyna Niemirowicz-Szczytt
Grzegorz Bartoszewski
Norikazu Tagashira
Yoshikazu Hoshi
Mark Borodovsky
Stanisław Karpiński
Stefan Malepszy
Zbigniew Przybecki