

CG020 Genomika

Bi7201 Základy genomiky

Přednáška 2

Identifikace genů

Jan Hejátko

Funkční genomika a proteomika rostlin,
Mendelovo centrum genomiky a proteomiky rostlin,
Středoevropský technologický institut (CEITEC), Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.muni.cz



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Literatura

▪ Zdrojová literatura ke kapitole 2

- Plant Functional Genomics, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey
- Majoros, W.H., Pertea, M., Antonescu, C. and Salzberg, S.L. (2003) GlimmerM, Exonomy, and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Research*, **31**(13).
- Singh, G. and Lykke-Andersen, J. (2003) New insights into the formation of active nonsense-mediated decay complexes. *TRENDS in Biochemical Sciences*, **28** (464).
- Wang, L. and Wessler, S.R. (1998) Inefficient reinitiation is responsible for upstream open reading frame-mediated translational repression of the maize R gene. *Plant Cell*, **10**, (1733)
- de Souza et al. (1998) Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins *PNAS*, **95**, (5094)
- Feuillet and Keller (2002) Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution *Ann Bot*, 89 (3-10)
- Frobis, A.C., Matus, D.Q., and Seaver, E.C. (2008). Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan *Capitella* sp. I. *PLoS One* 3, e4004

Osnova

- Postupy „přímé“ a reverzní genetiky
 - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
 - struktura genů a jejich vyhledávání
 - genomová kolinearita a genová homologie
- Experimentální identifikace genů
 - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
 - EST knihovny
 - přímá a reverzní genetika

Osnova

- Postupy „přímé“ a reverzní genetiky
 - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí

Přímá vs. reverzní genetika

Revoluce v chápání pojmu genu

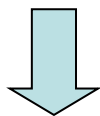
Přístupy „klasické“ genetiky



3

:

1

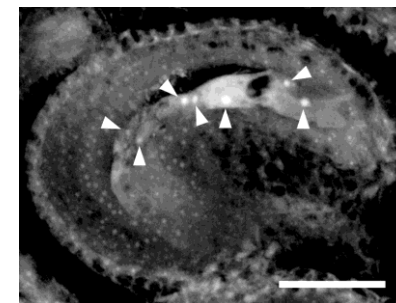
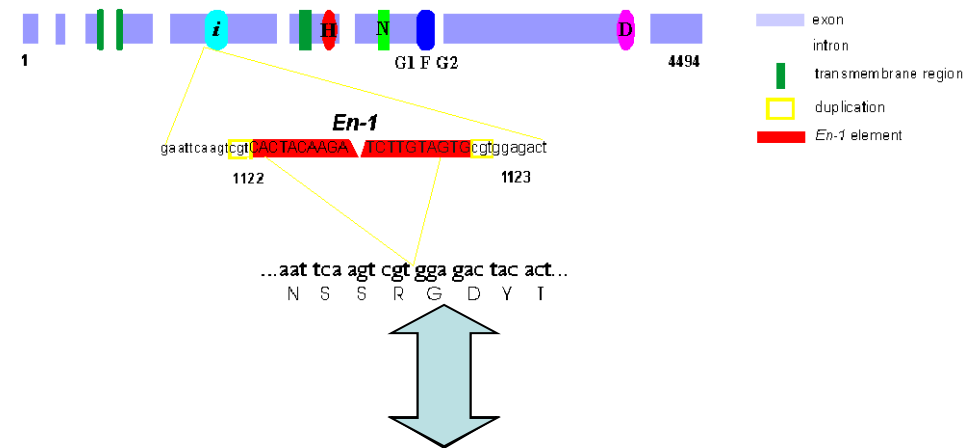
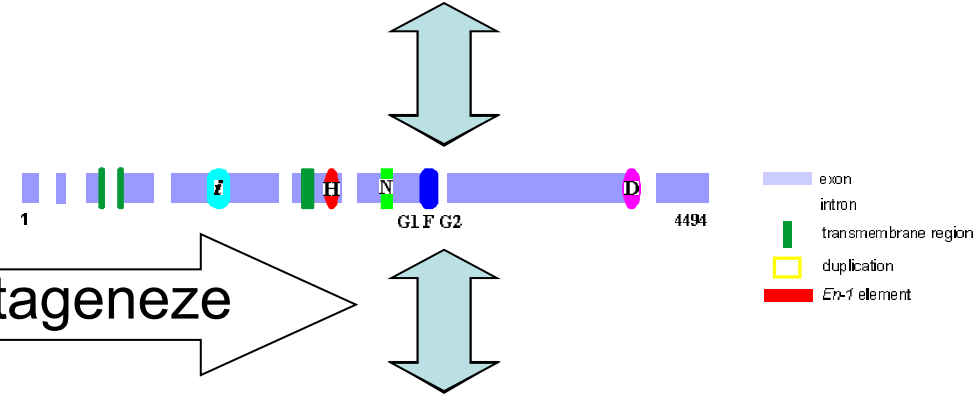


?



„Reverzně genetický“ přístup

5'TTATATATATATATTTAAAAAATAAAATAAAA
GAACAAAAAAGAAAATAAAATA....3'

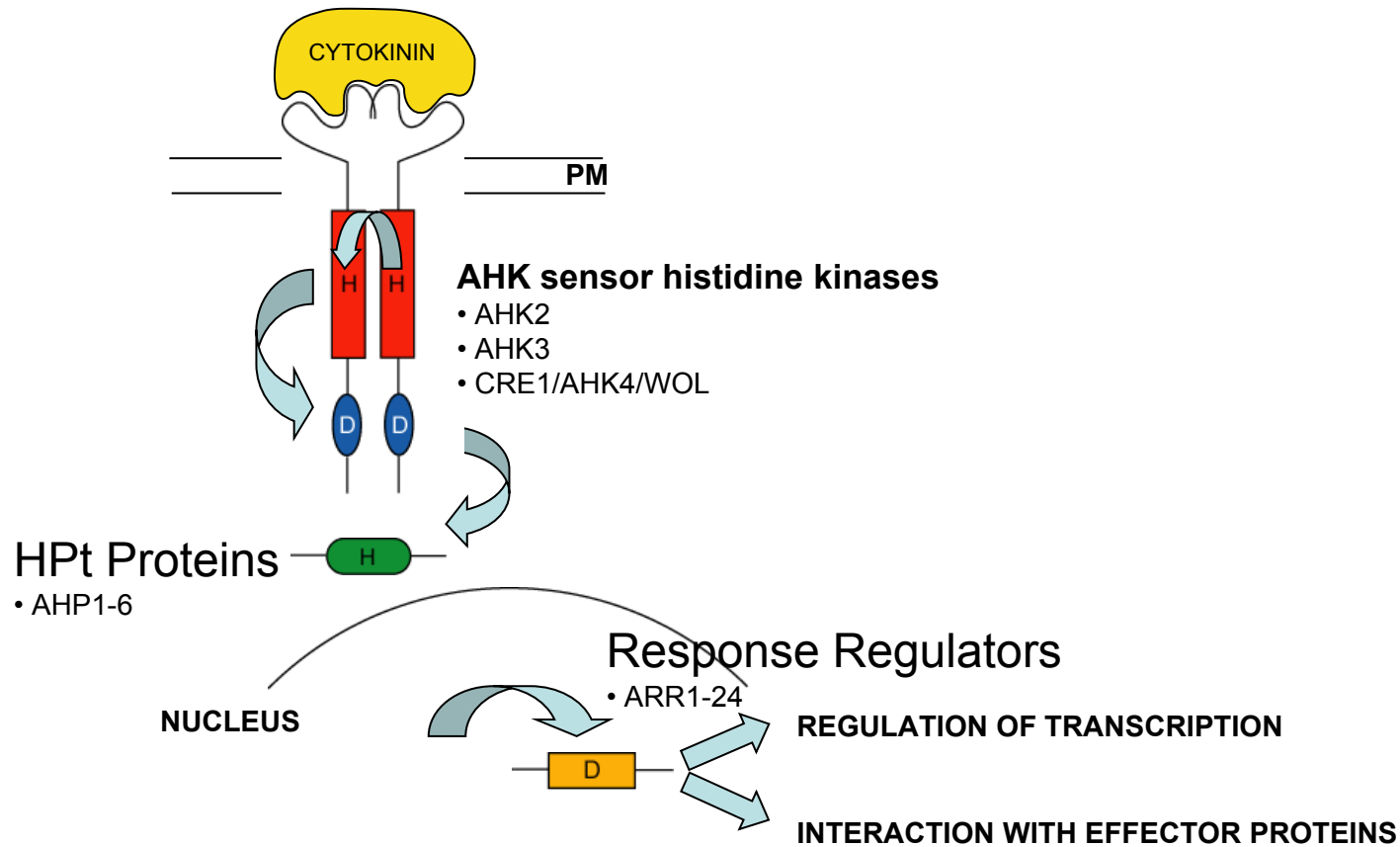


Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*

Identifikace role genu *ARR21*

Recent Model of the CK Signaling via Multistep Phosphorelay (MSP) Pathway



Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST

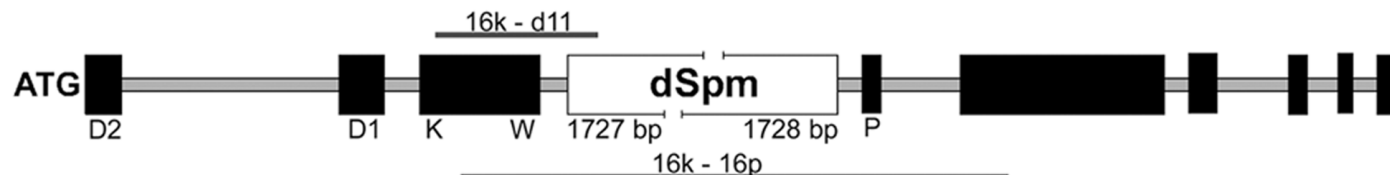
Identifikace role genu *ARR21* – izolace inz. mutanta

- vyhledávání v databázi inzerčních mutantů (SINS)

```
Insert_SINS: 01_09_64
Query: 80      tcctagcggttcatgagcgtaccatacttgacaanagagaacgtagccagccatttacagg 139
              |||
Sbjct: 58319  tcctagcggttcatgagcgtaccatacttgacaagagagaacgtagccagccatttacagg 58378
Arr21: 1830
```

```
Insert_SINS: 01_09_64
Query: 140     tttgatatctcttgtcaaaaatgtttttggattttactgt 179
              |||
Sbjct: 58379  tttgatatctcttgtcaaaaatgtttttggattttactgt 58418
Arr21: 1890
```

- lokalizace inserce *dSpm* v genomové sekvenci *ARR21* pomocí sekvenace PCR produktů



Identifikace role genu *ARR21*

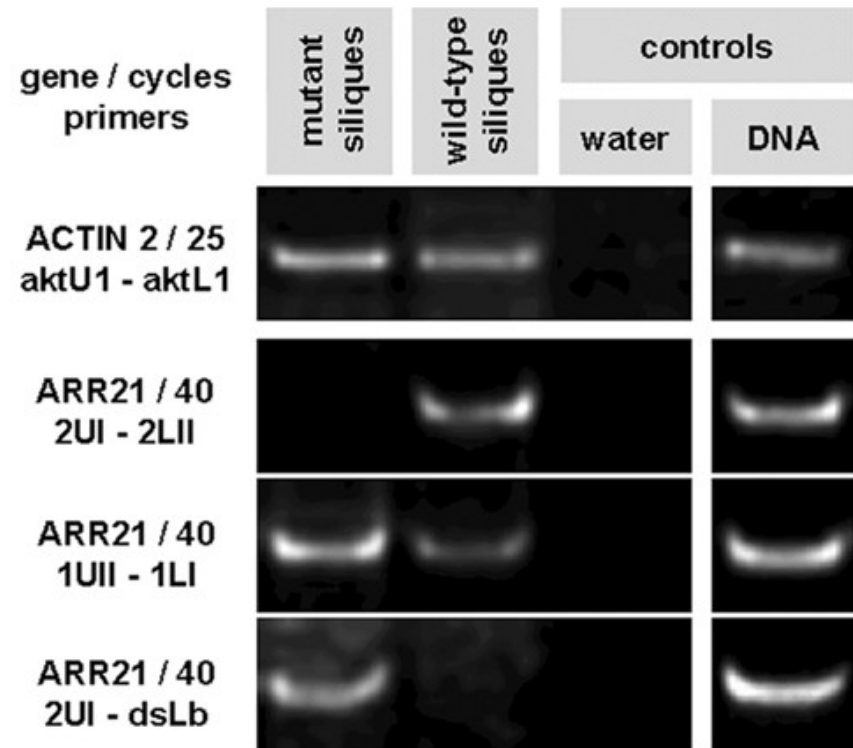
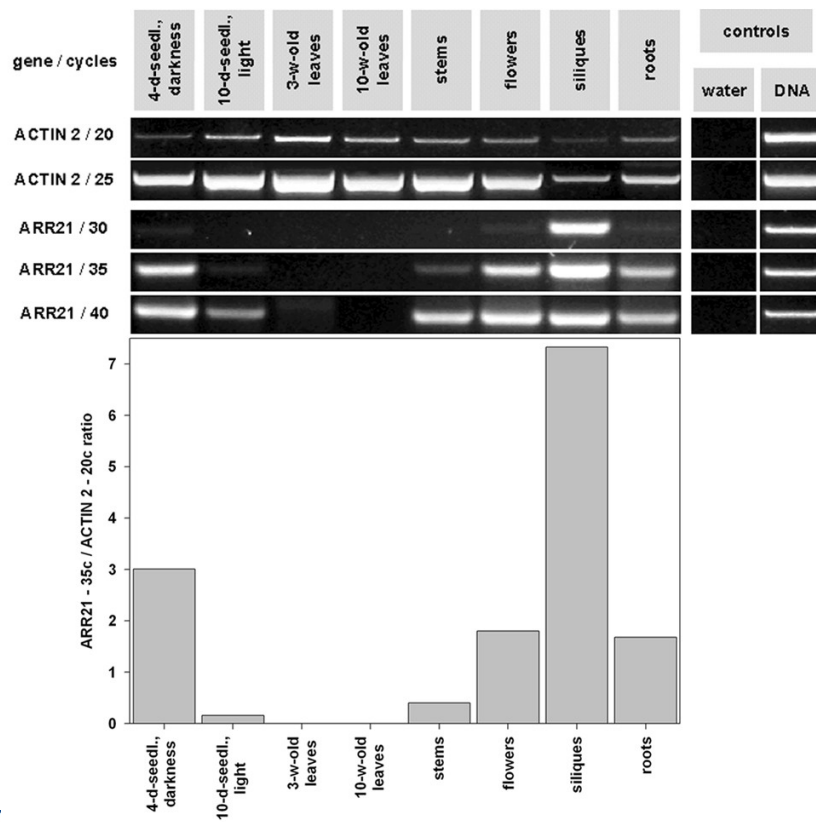
- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
- Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutanta potvrzena na úrovni RNA

Identifikace role genu

ARR21 – analýza exprese

Standardní typ

Inzerční mutant

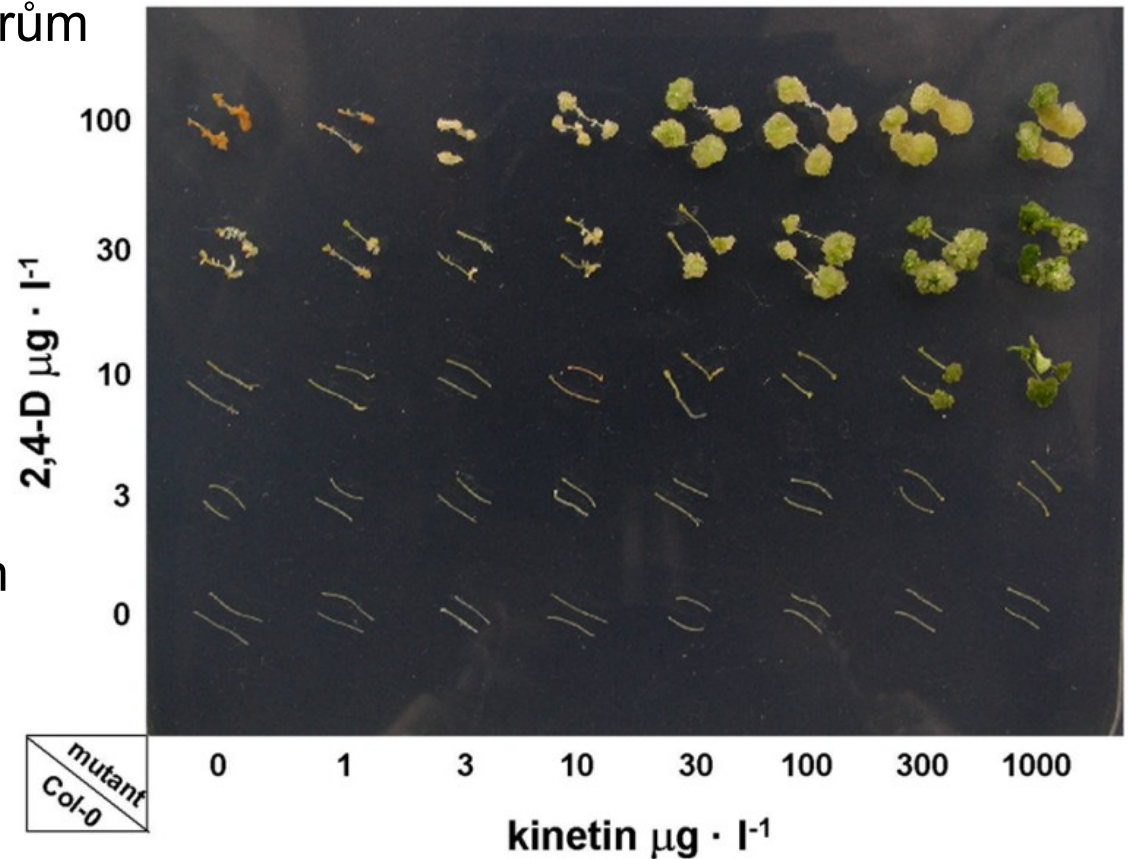


Identifikace role genu *ARR21*

- Předpokládaný přenašeč signálu u dvoukomponentního signálního systému *Arabidopsis*
- Mutant identifikován vyhledáváním v databázi inzerčních mutantů (SINS-sequenced insertion site) pomocí programu BLAST
- Exprese *ARR21* u standardního typu a Inhibice exprese u inzerčního mutanta potvrzena na úrovni RNA
- Analýza fenotypu inzerčního mutanta

Identifikace role genu *ARR21* – analýza fenotypu mutantu

- Analýza citlivosti k regulátorům růstu rostlin
 - 2,4-D a kinetin
 - etylén
 - světlo různých vlnových délek
- Doba kvetení i počet semen nezměněn



Identifikace role genu

ARR21 – příčiny absence fenotypu

- Funkční redundance v rámci genové rodiny?

Identifikace role genu

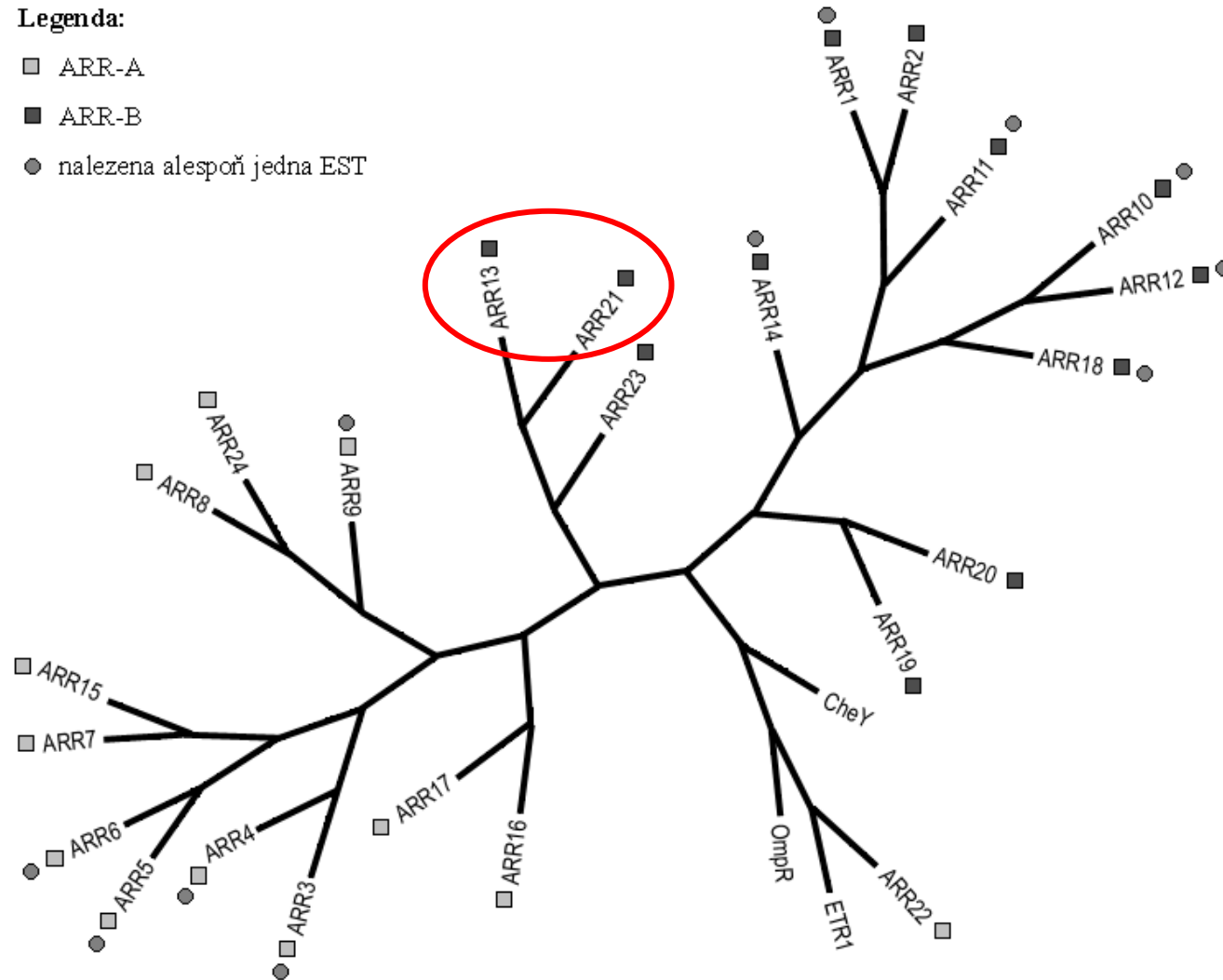
ARR21 – příbuznost ARR genů

Legenda:

□ ARR-A

■ ARR-B

● nalezena alespoň jedna EST



Identifikace role genu *ARR21* – příčiny absence fenotypu

- Funkční redundance v rámci genové rodiny?
- Fenotypový projev pouze za velmi specifických podmínek (?)

Identifikace role genu

ARR21 – shrnutí

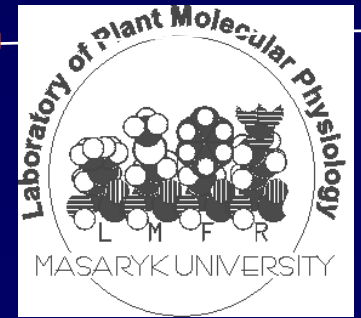
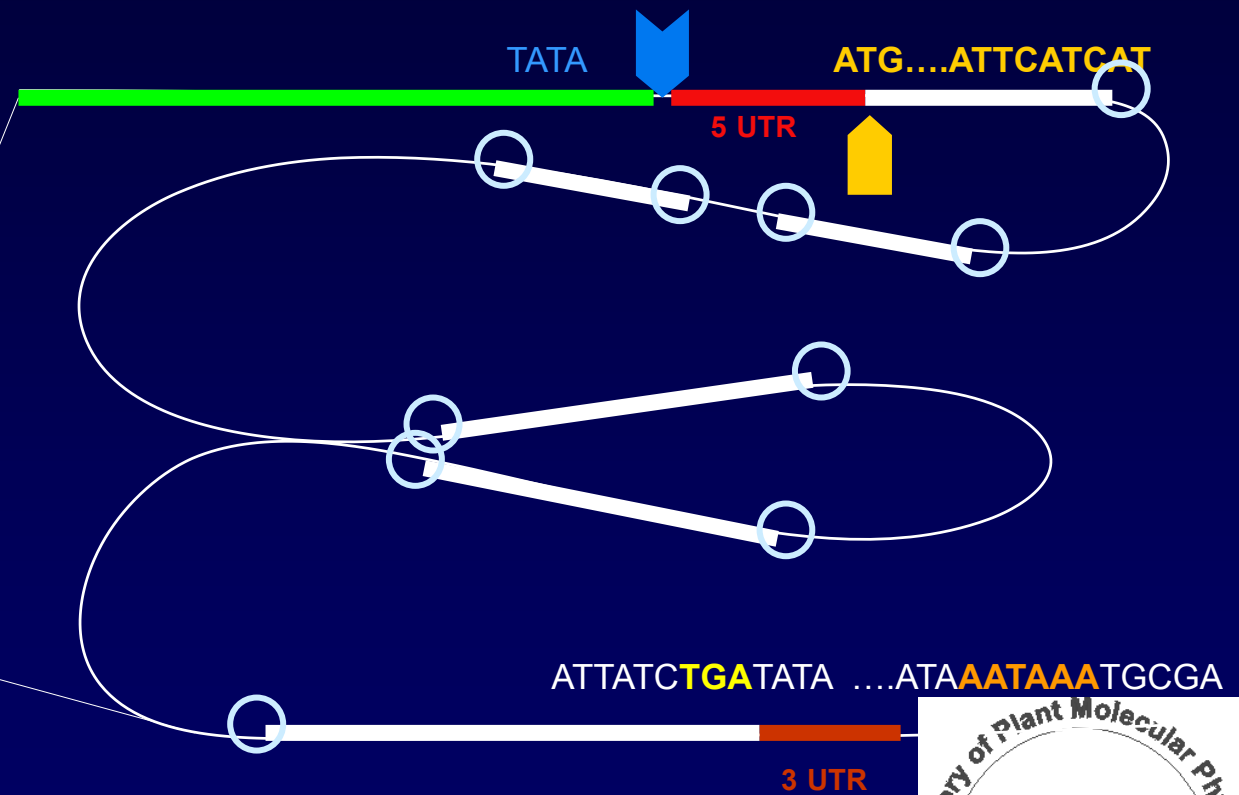
- Gen *ARR21* identifikován pomocí srovnávací analýzy genomu *Arabidopsis*
- Na základě analýzy sekvence byla předpovězena jeho funkce
- Byla prokázána místně specifická exprese genu *ARR21* na úrovni RNA
- Inzerční mutagenese v případě identifikace funkce genu *ARR21* ve vývoji *Arabidopsis* byla neúspěšná, pravděpodobně v důsledku funkční redundance v rámci genové rodiny

Osnova

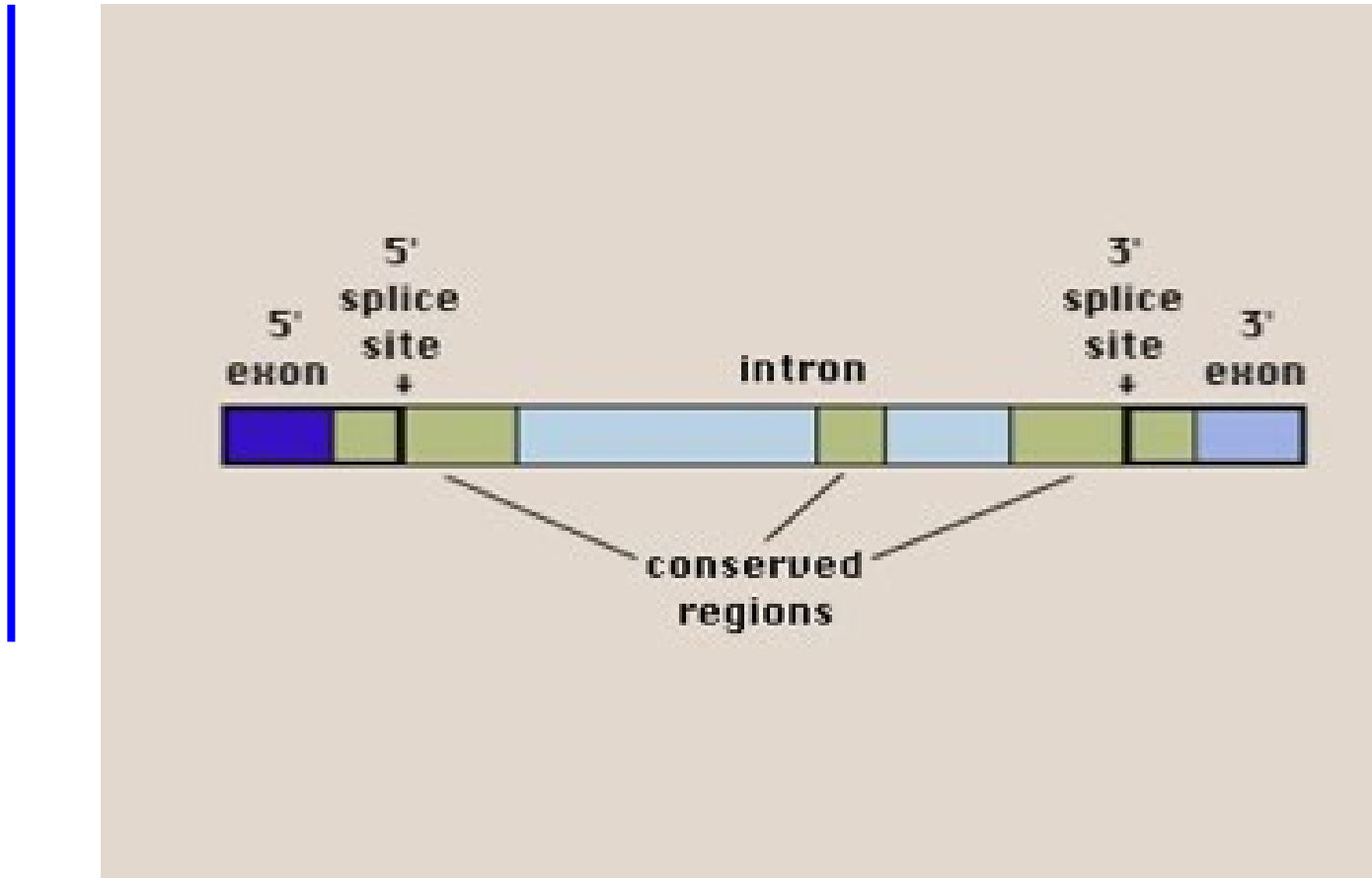
- Identifikace genů *ab initio*
 - struktura genů a jejich vyhledávání

Struktura genů

- promotor
- počátek transkripce
- 5 UTR
- počátek translace
- místa sestřihu
- stop kodon
- 3 UTR
- polyadenylační signál



Sestřih RNA



Identifikace genů *ab initio*

- zanedbání 5' a 3' UTR
- identifikace počátku translace (ATG) a stop kodonu (TAG, TAA, TGA)
- nalezení donorových (většinou GT) a akceptorových (AG) míst sestřihu
- většina ORF není skutečně kódujícími sekvencemi – u *Arabidopsis* je asi 350 mil. ORF na každých 900 bp (!)
- využití různých statistických modelů (např. Hidden Markov Model, HMM, viz doporučená studijní literatura, Majoros et al., 2003) k posouzení a ohodnocení váhy identifikovaných donorových a akceptorových míst

Predikce míst sestřihu

- programy pro predikci míst sestřihu (specificita přibližně 35%)
 - GeneSplicer (http://www.tigr.org/tdb/GeneSplicer/gene_spl.html)
 - SplicePredictor (<http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi>)

Predikce míst sestřihu

BCB @ ISU Bioinformatics 2 Download Help Tutorial References Contact
Go

SplicePredictor

- a method to identify potential splice sites in (plant) pre-mRNA by sequence inspection using Bayesian statistical models
(click [here](#) to access the older method using logitlinear models)

Sequences should be in the one-letter-code ({a,b,c,g,h,k,m,n,r,s,t,u,w,y}), upper or lower case; all other characters are ignored during input. Multiple sequence input is accepted in **FASTA** format (sequences separated by identifier lines of the form “>SQ;name_of_sequence comments”) or in **GenBank** format.

Paste your genomic DNA sequence here:

```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTTCGATCTCAGATATA  
AAAGATTTTCATTCAATATAAATACTTGGATAAATACTCTTATTATTTTTCTTTAGTTTATTAAAAAAAACCTCTAATAAAT  
ACGAGTTTTAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAAACGATAAAGTTTACAAAAGTAATATCC  
AAGTATCTCATAGTCAACATATATATAGTAATAATTAGTTGACGTATAAGAAAATAAAAAATAAATAAATTAGTATCTTAT  
TTTGGGTGGTGCTGACTGGTGACTGGTGACTGCAGAATGCTCGGCAAATGGAACCATATCCCAAGACATGGGTTTTAGAT
```

... or upload your sequence file (specify file name):

Browse...

... or type in the GenBank accession number of your sequence:



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Predikce míst sestřihu

What do the output columns mean?

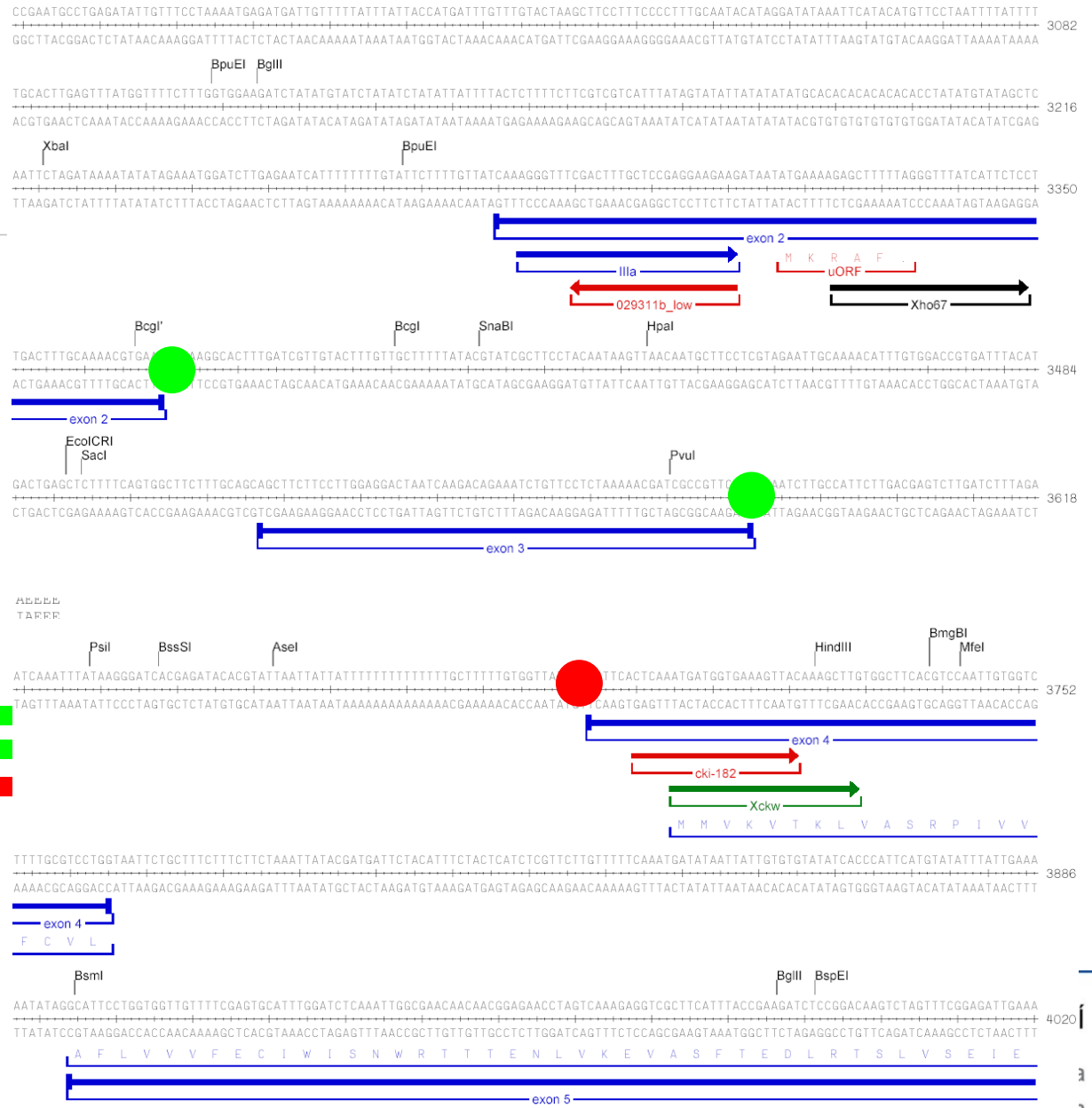
SplicePredictor. Version of February 13, 2005.
Date run: Wed Nov 9 11:30:14 2005

Species: Homo sapiens
Model: 2-class Bayesian
Prediction cutoff (2 ln[BF]): 3.00
Local pruning: on
Non-canonical sites: not scored

Sequence 1: your-sequence, from 1 to 9490.

Potential splice sites

t	q	loc	sequence	P	c	rho	gamma	*	P*R*G*
A	<--	75	ttttttcgatctcAGat	0.973	7.16	0.000	0.000	7	(5 1 1)
A	<--	134	attatttttctttAGtt	0.999	14.86	0.000	0.000	7	(5 1 1)
A	<--	500	gattttgtgtttAGtc	0.977	7.48	0.000	0.000	7	(5 1 1)
A	<--	780	tctgttattgtatAGct	0.986	8.56	0.000	0.000	7	(5 1 1)
A	<--	848	tattttttgaaatAGat	0.968	6.80	0.000	0.000	7	(5 1 1)
A	<--	1051	caatttatttttaAGaa	0.930	5.19	0.000	0.000	7	(5 1 1)
A	<--	1213	ttatttattttttAGtt	0.998	12.14	0.000	0.000	7	(5 1 1)
A	<--	1373	tttctctctcacAGga	0.999	13.17	0.000	0.000	7	(5 1 1)
A	<--	1487	tttataatttgatAGtg	0.883	4.04	0.000	0.000	7	(5 1 1)
A	<--	1581	atgtgttctgttAGga	0.982	8.03	0.000	0.000	7	(5 1 1)
A	<--	1781	ggttgtgcaaatAGgg	0.886	4.10	0.000	0.000	7	(5 1 1)
A	<--	2440	taataaaaatttAGat	0.939	5.46	0.000	0.000	7	(5 1 1)
A	<--	2479	catctaaaatttAGat	0.942	5.59	0.000	0.000	7	(5 1 1)
D	---->	2546	aagGTtagta	0.909	4.61	0.885	1.903	15	(5 5 5)
A	<--	2572	tttttttttggcAGca	0.930	5.16	0.000	0.000	7	(5 1 1)
A	<----	2763	ctcaaatccaaaAGgt	0.873	3.86	0.185	0.000	11	(5 5 1)
A	<----	2782	tttcgtttccattAGcg	0.952	5.98	0.220	0.000	11	(5 5 1)
A	<----	3022	tttgtttgtaactAGct	0.956	6.16	0.221	0.000	11	(5 5 1)
A	<----	3048	ctttgcaattacatAGga	0.973	7.15	0.229	0.000	11	(5 5 1)
A	<--	3171	cgctgctatttatAGta	0.988	8.74	0.000	0.000	7	(5 1 1)
A	<--	3284	cttttttttcaaaAGgg	0.993	10.03	0.000	0.006	8	(5 1 2)
A	<----	3451	aatgcttctctgAGaa	0.916	4.77	0.293	0.065	12	(5 5 2)
D	---->	3649	cacGTatta	0.933	5.25	0.000	1.848	11	(5 1 5)
A	<--	4254	attattgttcttcAGat	0.998	12.82	0.000	0.002	8	(5 1 2)
A	<--	4351	tttcttacattgcAGaa	0.991	9.42	0.000	0.000	7	(5 1 1)
A	<--	4633	gtcttgttctcttAGgg	0.879	3.97	0.000	0.000	7	(5 1 1)
A	<--	4976	cttgtgtttctcAGct	0.952	5.98	0.000	0.000	7	(5 1 1)
A	<--	5004	tttttttttggcAGag	0.996	11.17	0.000	0.000	7	(5 1 1)
D	---->	5356	caagGTaat	0.821	3.04	0.387	0.000	11	(5 5 1)
D	---->	5384	ttgGTaaga	0.941	5.54	0.478	0.090	13	(5 5 3)
A	<--	5403	actctgtttcttAGct	0.894	4.26	0.000	0.000	7	(5 1 1)
A	<----	5441	ctttctcttaacAGaa	0.995	10.43	0.387	0.000	11	(5 5 1)
A	<--	5472	ttgttaaaattacAGct	0.965	6.62	0.478	0.090	13	(5 5 3)
D	---->	5745	gcgGTaaga	0.991	9.48	0.990	1.956	15	(5 5 5)
A	<----	5808	catcatatcctaaAGgt	0.948	5.83	0.458	0.000	11	(5 5 1)
A	<----	6135	ggtctatttattAGgt	0.999	13.59	0.508	0.050	12	(5 5 2)
A	<--	6552	ggattttcacctcAGag	0.938	5.42	0.000	0.000	7	(5 1 1)



Identifikace genů *ab initio*

- programy pro predikci míst sestřihu (specifická přibližně 35%)
 - GeneSplicer (http://www.tigr.org/tdb/GeneSplicer/gene_spl.html)
 - SplicePredictor (<http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi>)
 - NetGene2 (<http://www.cbs.dtu.dk/services/NetGene2/>)

Predikce míst sestřihu



[CBS](#) >> [Prediction Servers](#) >> [NetGene2](#)

NetGene2 Server

The NetGene2 server is a service producing neural network predictions of splice sites in human, *C. elegans* and *A. thaliana*

[Instructions](#)

[Output format](#)

[Abstract](#)

[Performance](#)

SUBMISSION

Submission of a local file with a single sequence:

File in **FASTA** format

- Human
 C. elegans
 A. thaliana

Submission by pasting a single sequence:

Sequence name

- Human
 C. elegans
 A. thaliana

Sequence

```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTTCGATC
TCAGATATA
AAAGATTTTCATTCAATATAATACTTGGATAAACTACTCTTATTATTTTTCTTTAGTTTATTAACAAAAACCT
CTAATAAAT
ACGAGTTTAAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTCAAACGATAAAGTTTACAAAA
```

NOTE: The submitted sequences are kept confidential and will be erased immediately after processing.

Predikce míst sestřihu

Prediction done

***** NetGene2 v. 2.4 *****

The sequence: Sequence has the following composition:

Length: 9490 nucleotides.

31.8% A, 17.0% C, 19.6% G, 31.7% T, 0.0% X, 36.5% G+C

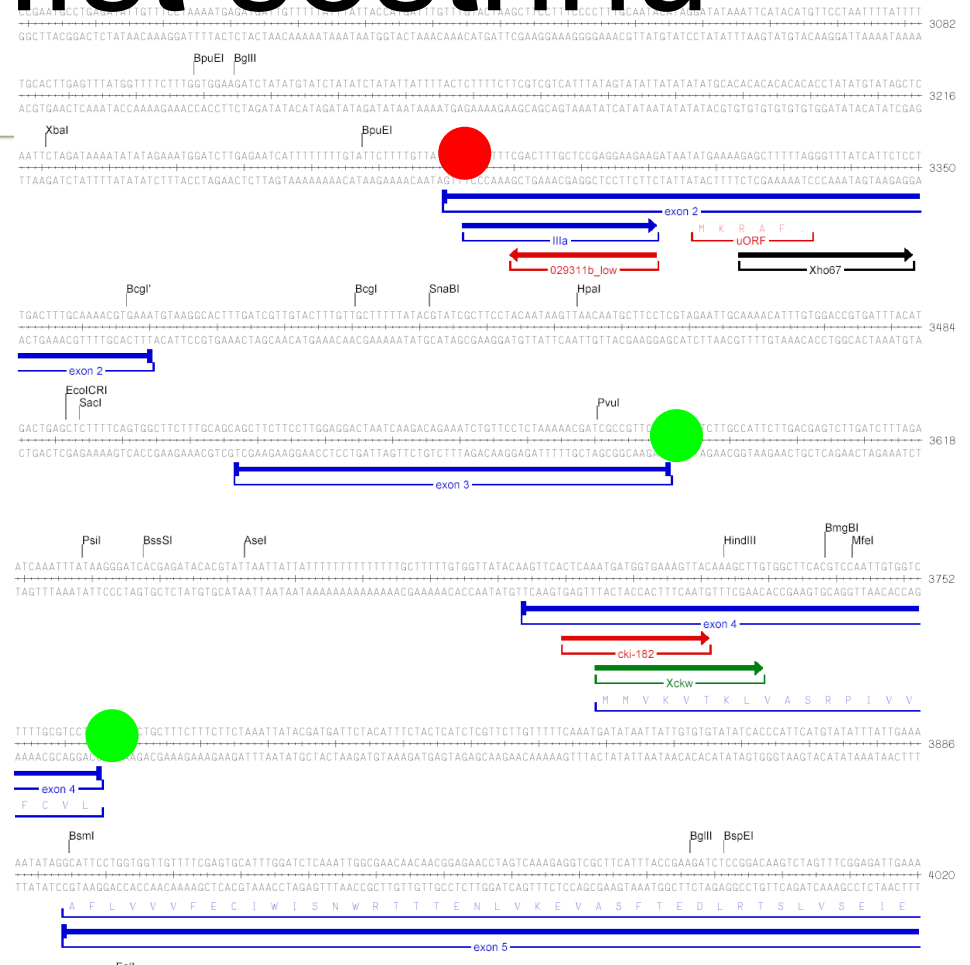
Donor splice sites, direct strand

pos 5'->3'	phase	strand	confidence	5'	exon	intron	3'
1704	0	+	0.87	TTC	CAACAC	^	TAATATTT
1906	0	+	0.99	CGG	TAACGG	^	CAGAACAT
4134	0	+	0.74	TCA	AACACAG	^	TGTTAAAA
4619	1	+	0.74	AGC	AAGAAAG	^	CTTGTTTC
4915	0	+	0.94	CGT	TCTCTG	^	AAATACTG
5356	0	+	0.87	TCT	CAACCA	^	GAATGTTT
5384	1	+	1.00	GAT	TGTTG	^	AAGACTCT
5809	1	+	1.00	TAT	CCTAAAG	^	STGTCCAA
6057	0	+	1.00	GC	AGTCTT	^	AAGCTACT
6096	1	+	0.74	CTC	TTCACAA	^	AAATCTAG
7369	0	+	1.00	GG	ACTGCCAA	^	AAGTTTAA
7886	0	+	0.74	GA	ACAAAATG	^	TAGATGAA
9323	0	+	0.74	GA	AGATTAGG	^	TTTTCTCT

Donor splice sites, complement strand

pos 3'->5'	pos 5'->3'	phase	strand	confidence	5'	intron	exon	3'
1213	0	+	0.59	TAT	TTTTT	^	TATGGAGAC	
1221	2	+	0.87	AGT	TATGG	^	ACAAGAATCG	
1373	0	+	0.71	TCT	CTCAC	^	GACACAGAAT	
1487	1	+	0.81	AT	TTGAT	^	TGGGACATTA	
4254	0	+	1.00	TGT	TCTTC	^	ATCGCACCAT	
4832	2	+	0.54	AAA	ATTGC	^	TTCAGTGGC	
5004	0	+	0.94	TTT	TGCCC	^	AGATACACAC	
5472	1	+	0.96	AAA	ATTAC	^	CTCTGCTCAA	
6135	0	+	1.00	ATT	ATTAT	^	GTAAGATTAA	
6490	1	+	0.90	AA	AGTTAC	^	TGGTGGAGAA	
6744	0	+	0.59	TG	TCAAAC	^	TTTCGTAGAG	
7447	0	+	0.96	TT	CTGCAC	^	ATGCCAGAAA	
7780	2	+	0.76	TCC	ATTTC	^	ATACAGAACA	
7786	2	+	0.92	TC	GATAC	^	AACACATGCA	

Acceptor splice sites, direct strand



Sestřih RNA a adaptace

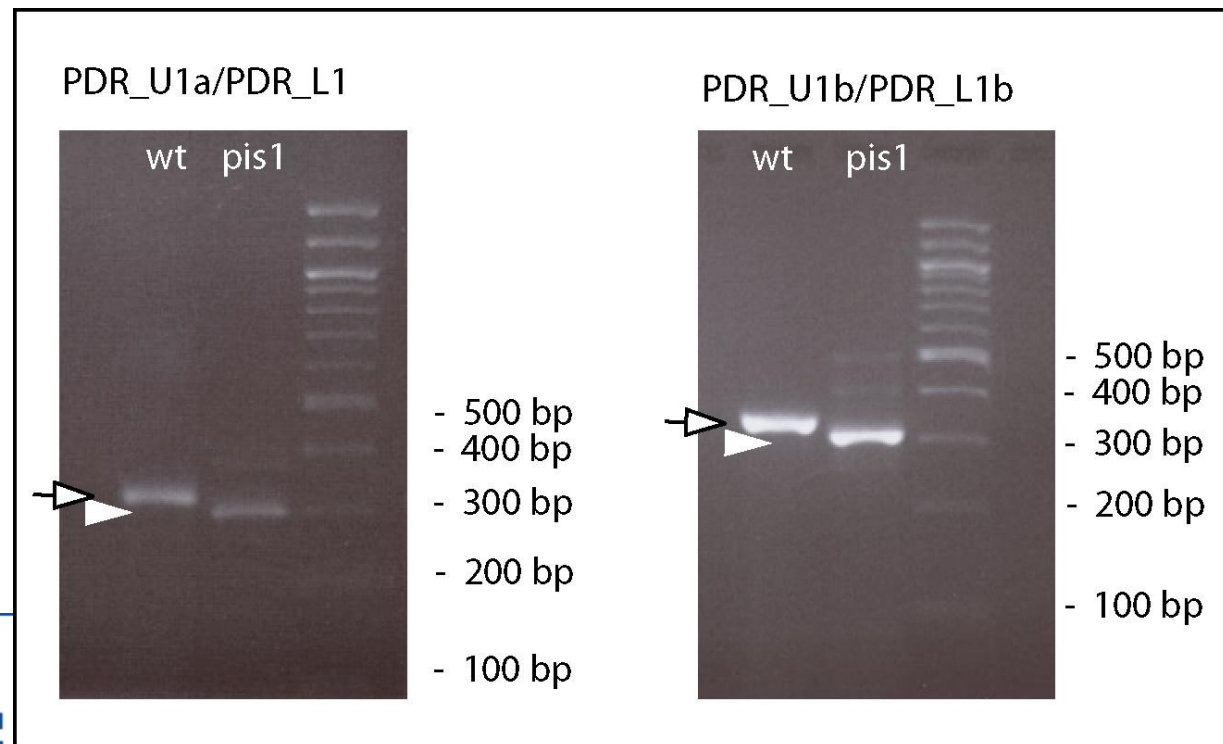
- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin

- identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu



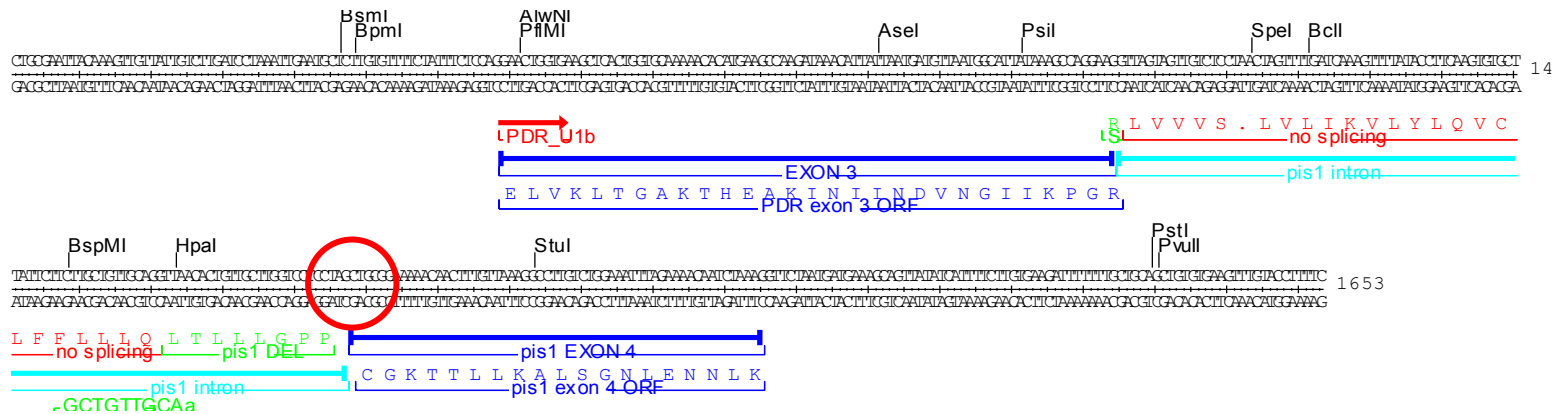
Sestřih RNA a adaptace

- identifikace mutantu s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
- analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu



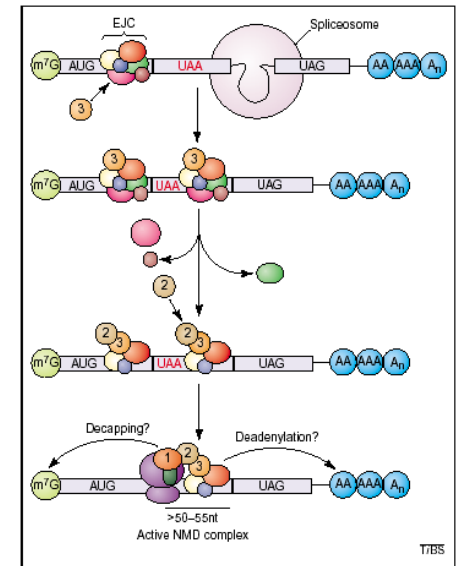
Identifikace genů *ab initio*

- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin
 - identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
 - analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu
 - sekvenace tohoto fragmentu pak ukázala na alternativní sesřih s využitím nejbližšího možného místa sestřihu v exonu 4



Identifikace genů *ab initio*

- odchylky rozpoznávání míst sestřihu u rostlin v praxi - příklad vývojové plasticity (nejen) rostlin
 - identifikace mutanta s bodovou mutací (tranzice G→A) přesně v místě sestřihu na 5' konci 4. exonu
 - analýza pomocí RT PCR prokázala přítomnost fragmentu kratšího než by odpovídalo cDNA po normálním sestřihu
 - sekvenace tohoto fragmentu pak ukázala na alternativní sestřih s využitím nejbližšího možného místa sestřihu v exonu 4
 - existence podobných obranných mechanismů prokázána i u jiných organismů (např. nestabilita mutantní mRNA se vznikem předčasného stopkodonu (> 50-55 bp před normálním stop kodonem) u eukaryot, viz doporučená studijní literatura, Singh and Lykke-Andersen, 2003)




Identifikace genů *ab initio*


- programy pro predikci exonů
 - 4 typy exonů (podle polohy):
 - iniciační
 - vnitřní
 - terminální
 - jednoduché
 - programy kromě rozpoznávání míst sestřihu zohledňují i strukturu jednotlivých typů exonů
- iniciační:
 - Genescan (<http://genes.mit.edu/GENSCAN.html>)
 - GeneMark.hmm (<http://opal.biology.gatech.edu/GeneMark/>)
- interní:
 - MZEF (<http://rulai.cshl.org/tools/genefinder/>)

Identifikace genů *ab initio*

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA



 [For information about Genscan, click here](#)

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences to process, request a local copy of the program (see instructions at the bottom of this page) or use the [GENSCAN email server](#). If your browser, e.g., Lynx does not support file upload or multipart forms, use the [older version](#).

Organism: Suboptimal exon cutoff (optional):

Sequence name (optional):

Print options: Predicted peptides only

Upload your DNA sequence file (one-letter code, upper or lower case, spaces/numbers ignored):

Or paste your DNA sequence here (one-letter code, upper or lower case, spaces/numbers ignored):

```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTCGATC
TCAGATATA
AAAGATTTTCATTCAATATAACTTGGATAAATACTCTTATTATTTTCTTTAGTTTATAAAAAACCT
CTAATAAAT
ACGAGTTTAAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAACAGATAAAGTTTACAAA
GTAATATCC
AAGTATCTCATAGTCAACATATATATAGTAATAATTAGTTGACGTATRAAGAAAATAAAAAATAAATTA
GTATCTTAT
TTTGGGTGGTCTGACTGGTACTGGTACTGCAGAATGCTCGGCAAAATGGAACCATATCCCAAGACATGG
GTTTTAGAT
AGAACAAAATAAGTGTCCGAAGGAATGATATTTAAAAGTCAAATAGAATAATTATAAATATTGTAATTAGCA
AATAAAAAC
```

To have the results mailed to you, enter your email address here (optional):

[Back to the top](#)

Identifikace genů *ab initio*

GENSCANW output for sequence CKII

GENSCAN 1.0 Date run: 10-Nov-105 Time: 02:24:26

Sequence CKII : 9490 bp : 36.53% C+G : Isochore 1 (0 - 43 C+G%)

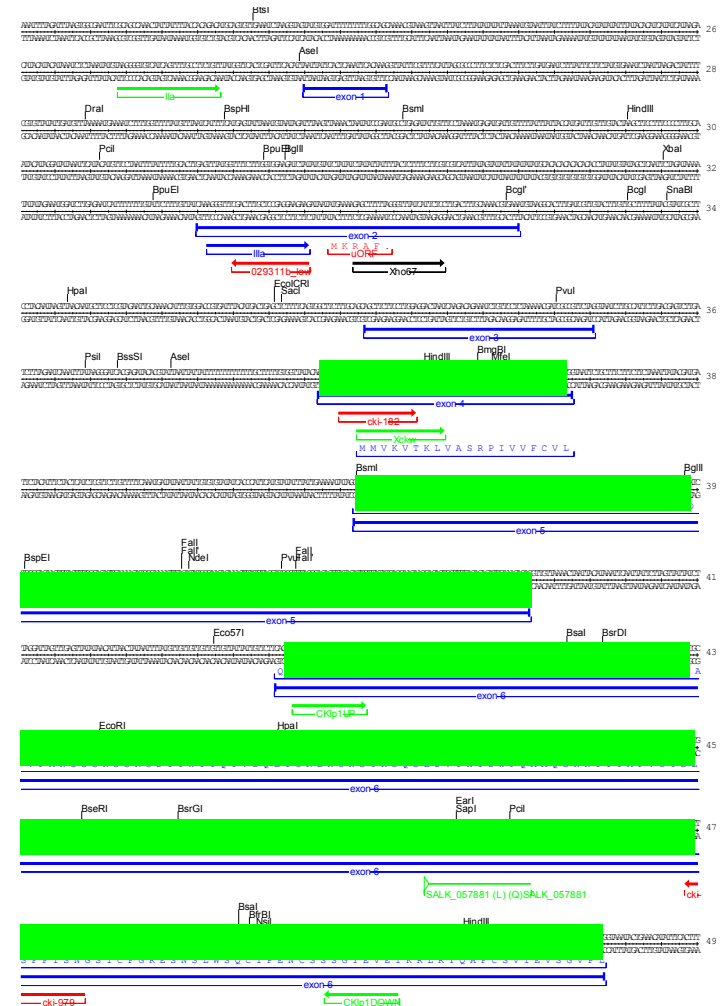
Parameter matrix: Arabidopsis.smat

Predicted genes/exons:

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.00	Prom	+	1497	1536	40							-3.85
1.04	Intr	+	5005	5383	379	0	1	70	91	343	0.772	31.41
1.05	Intr	+	5473	6056	584	2	2	38	99	582	0.722	50.76
1.06	Intr	+	6136	7368	1233	0	0	68	108	655	0.977	56.86
1.07	Term	+	7448	7660	213	1	0	43	35	212	0.999	12.65
1.08	PlyA	+	7910	7915	6							-0.45
2.03	PlyA	-	7976	7971	6							-4.83
2.02	Term	-	8793	8050	744	0	0	107	37	542	0.997	48.46
2.01	Init	-	9253	8936	318	1	0	105	73	386	0.999	41.18

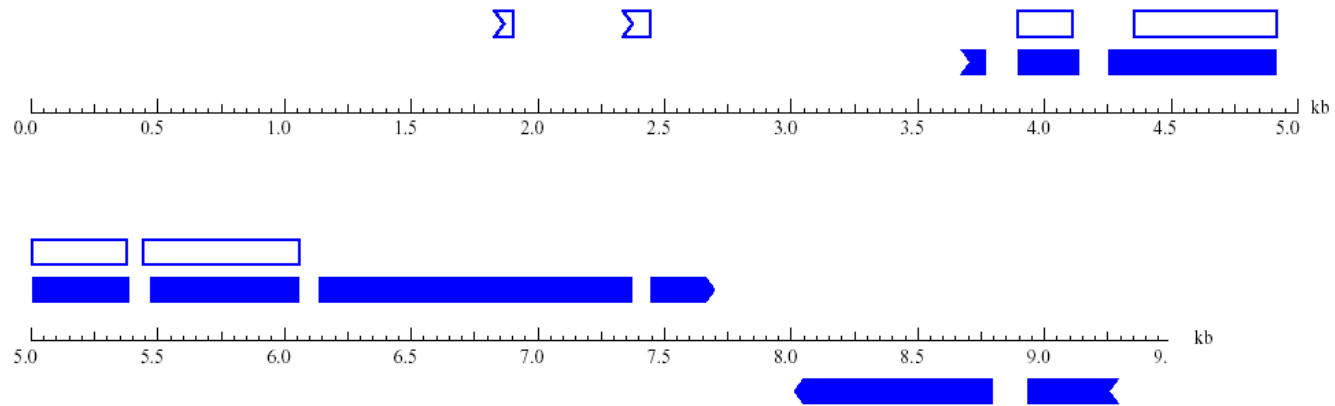
Suboptimal exons with probability > 0.100

Exnum	Type	S	.Begin	...End	.Len	Fr	Ph	B/Ac	Do/T	CodRg	P....	Tscr..
S.001	Init	+	1867	1905	39	0	0	64	40	57	0.298	3.74
S.002	Init	+	2374	2442	69	0	0	55	95	-11	0.132	2.40
S.003	Intr	+	3894	4110	217	2	1	-3	-34	307	0.177	11.55
S.004	Intr	+	4352	4914	563	0	2	75	59	338	0.187	26.20
S.005	Intr	+	5005	5379	375	0	0	70	8	335	0.212	22.99
S.006	Intr	+	5442	6056	615	2	0	95	99	589	0.208	57.32



Identifikace genů *ab initio*

GENSCAN predicted genes in sequence 02:56:23



Key:



Initial
exon



Internal
exon



Terminal
exon



Single-exon
gene



Optimal exon

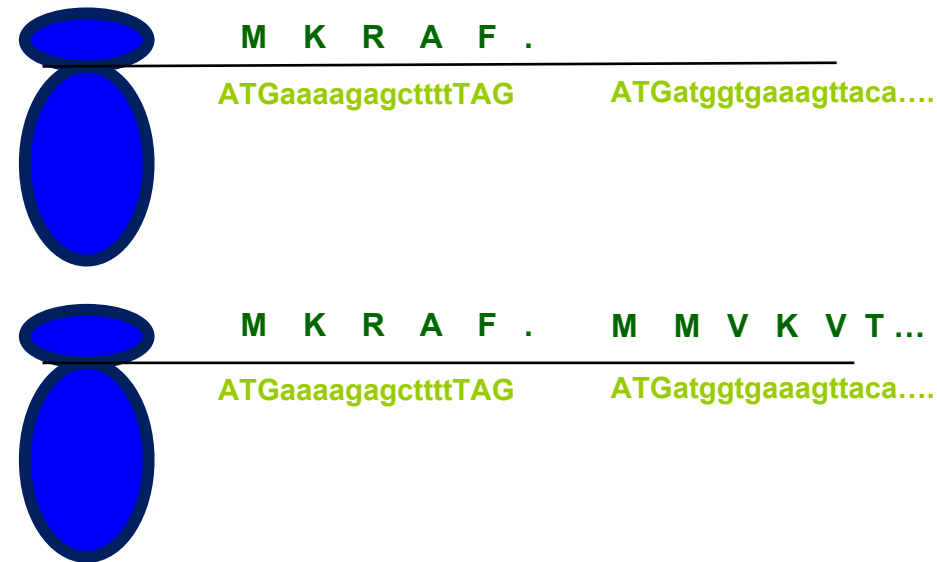


Suboptimal exon

Regulace transkripce

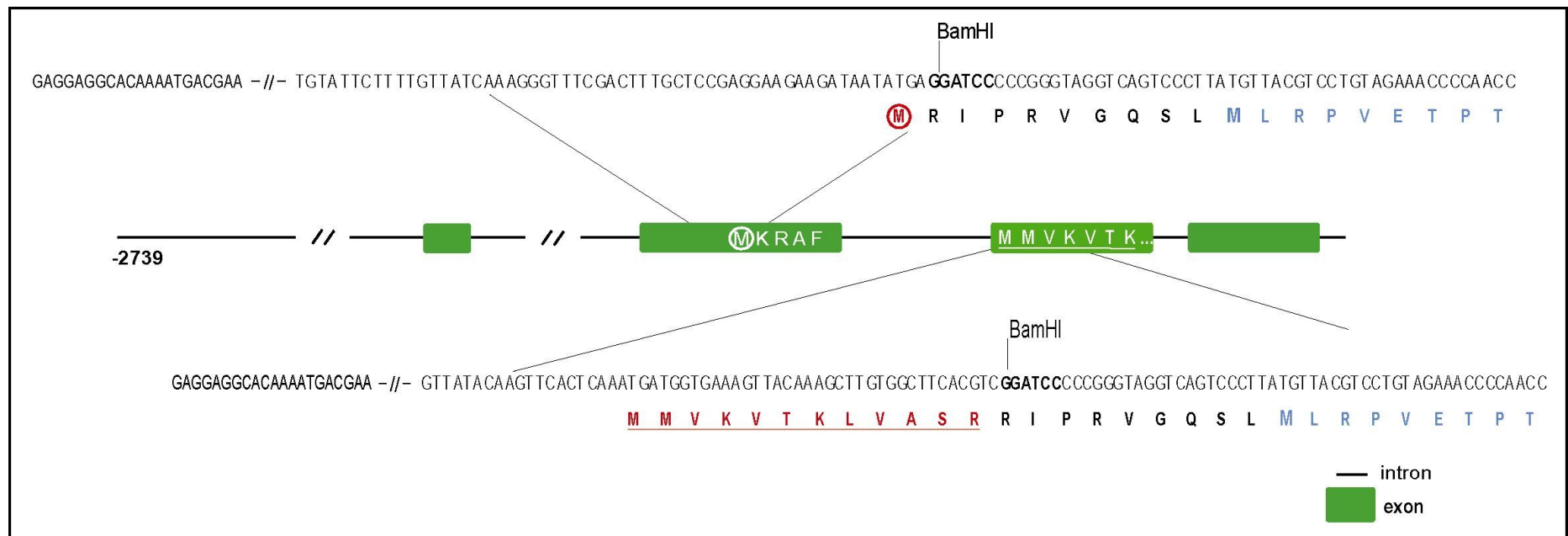
- Funkční význam sestřihu v nepřekládaných oblastech - důležitá regulační součást genů

- Translační represe prostřednictvím krátkých ORF v 5'UTR
- Identifikováno např. u kukuřice (Wang and Wessler, 1998, viz doporučená lit.)
- V případě CKI1 pokus prokázat tento způsob regulace genové exprese pomocí transgenních linií nesoucích *uidA* pod kontrolou dvou verzí promotoru, zatím nepotvrzeno



Regulace transkripce

- Funkční význam sestřihu v nepřekládaných oblastech - důležitá regulační součást genů
- V případě CKI1 pokus prokázat tento způsob regulace genové exprese pomocí transgenních linií nesoucích *uidA* pod kontrolou dvou verzí promotoru, zatím nepotvrzeno



Genové modelování

- programy pro genové modelování
 - zohledňují také další parametry, např. návaznost ORF
 - Genescan (<http://genes.mit.edu/GENSCAN.html>)
velice dobrý pro predikci exonů v kódujících oblastech
(testováno na genu *PDR9*, identifikoval všech 23 (!) exonů)
 - GeneMark.hmm (<http://opal.biology.gatech.edu/GeneMark/>)

Identifikace genů *ab initio*

Result of last submission:

[View PDF Graphical Output](#)

GeneMark.hmm Listing

Go to: [GeneMark.hmm Protein Translations](#)

Go to: [Job Submission](#)

Eukariotic GeneMark.hmm version bp 3.9 April 25, 2008

Sequence name: CK11

Sequence length: 5043 bp

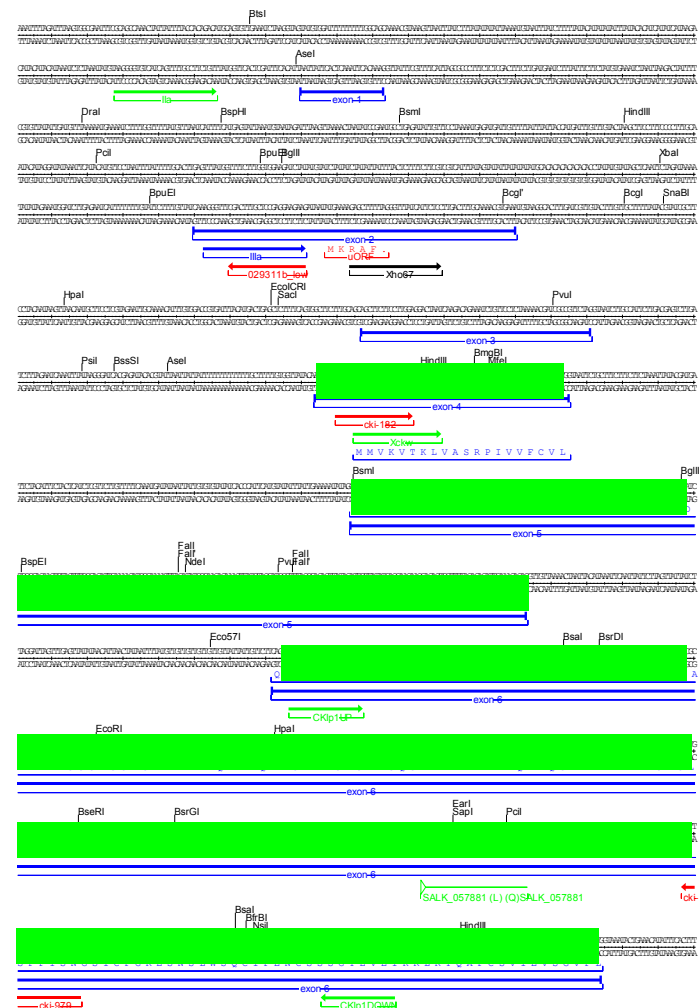
G+C content: 38.79%

Matrices file: /home/genemark/euk_gbm.matrices/mthaliana_hmm3.0mod

Thu Oct 1 11:09:24 2009

Predicted genes/exons

Gene #	Exon #	Strand	Exon Type	Exon Range	Exon Length	Start/End Frame
1	1					1 3 - -
1	2					1 3 - -
1	4	+	Internal	2266 - 2644	379	1 1 - -
1	5	+	Internal	2734 - 3317	584	2 3 - -
1	6	+	Internal	3397 - 4629	1233	1 3 - -
1	7	+	Terminal	4709 - 4921	213	1 3 - -



VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Identifikace genů *ab initio*

Result of last submission:

[View PDF Graphical Output](#)

GeneMark.hmm Listing

Go to: [GeneMark.hmm Protein Translations](#)

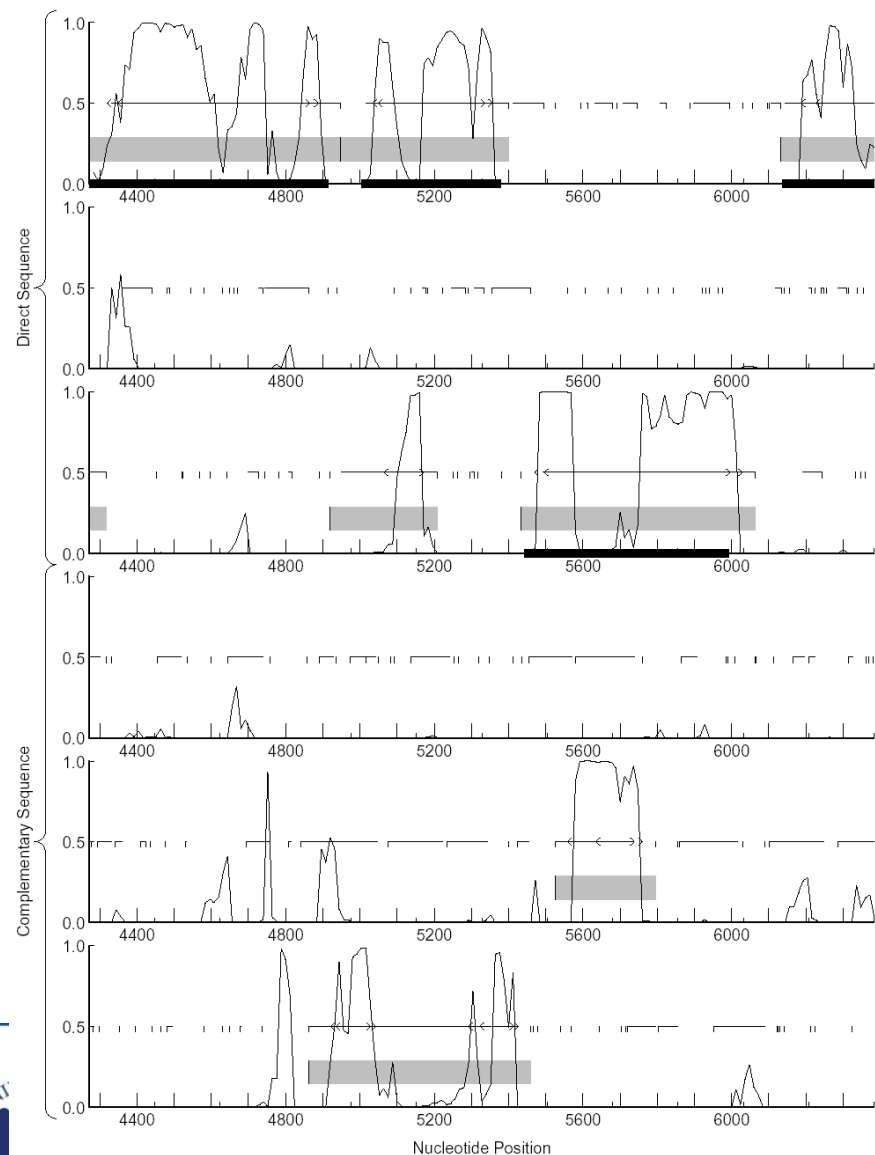
Go to: [Job Submission](#)

Eukariotyc GeneMark.hmm version bp 3.9 April 25, 2008
 Sequence name: CK11
 Sequence length: 5043 bp
 G+C content: 38.79%
 Matrices file: /home/genemark/euk_gbm.matrices/athaliana_hmm3.0mod
 Thu Oct 1 11:09:24 2009

Predicted genes/exons

Gene #	Exon #	Strand	Exon Type	Exon Range	Exon Length	Start/End Frame
1	1	+	Initial	969 1025 57 1 3 - -		
1	2	+	Internal	1155 1394 240		1 3 - -
1	3	+	Internal	1516 2175 660		1 3 - -
1	4	+	Internal	2266 2644 379		1 1 - -
1	5	+	Internal	2734 3317 584		2 3 - -
1	6	+	Internal	3397 4629 1233		1 3 - -
1	7	+	Terminal	4709 4921 213		1 3 - -

GeneMark.hmm prediction Thu Nov 10 03:23:47 EST 2005, Order 5, Window 96, Step 12, 4/6



...ÁVÁNÍ

ancována

Evropským sociálním fondem
a státním rozpočtem České republiky

Genové homologie

- vyhledávání genů podle homologií
 - porovnávání s EST databázemi
 - **BLASTN** (<http://www.ncbi.nlm.nih.gov/BLAST/>, <http://workbench.sdsc.edu/>)
 - porovnávání s proteinovými databázemi
 - **BLASTX** (<http://www.ncbi.nlm.nih.gov/BLAST/>, <http://workbench.sdsc.edu/>)
 - **Genewise** (<http://www.ebi.ac.uk/Wise2/>)

porovnávají proteinovou sekvenci s genomovou DNA (po zpětném překladu), je nutná znalost aminokyselinové sekvence
 - porovnávání s homologními genomovými sekvencemi z příbuzných druhů
 - **VISTA/AVID** (<http://www.lbl.gov/Tech-Transfer/techs/lbnl1690.html>)

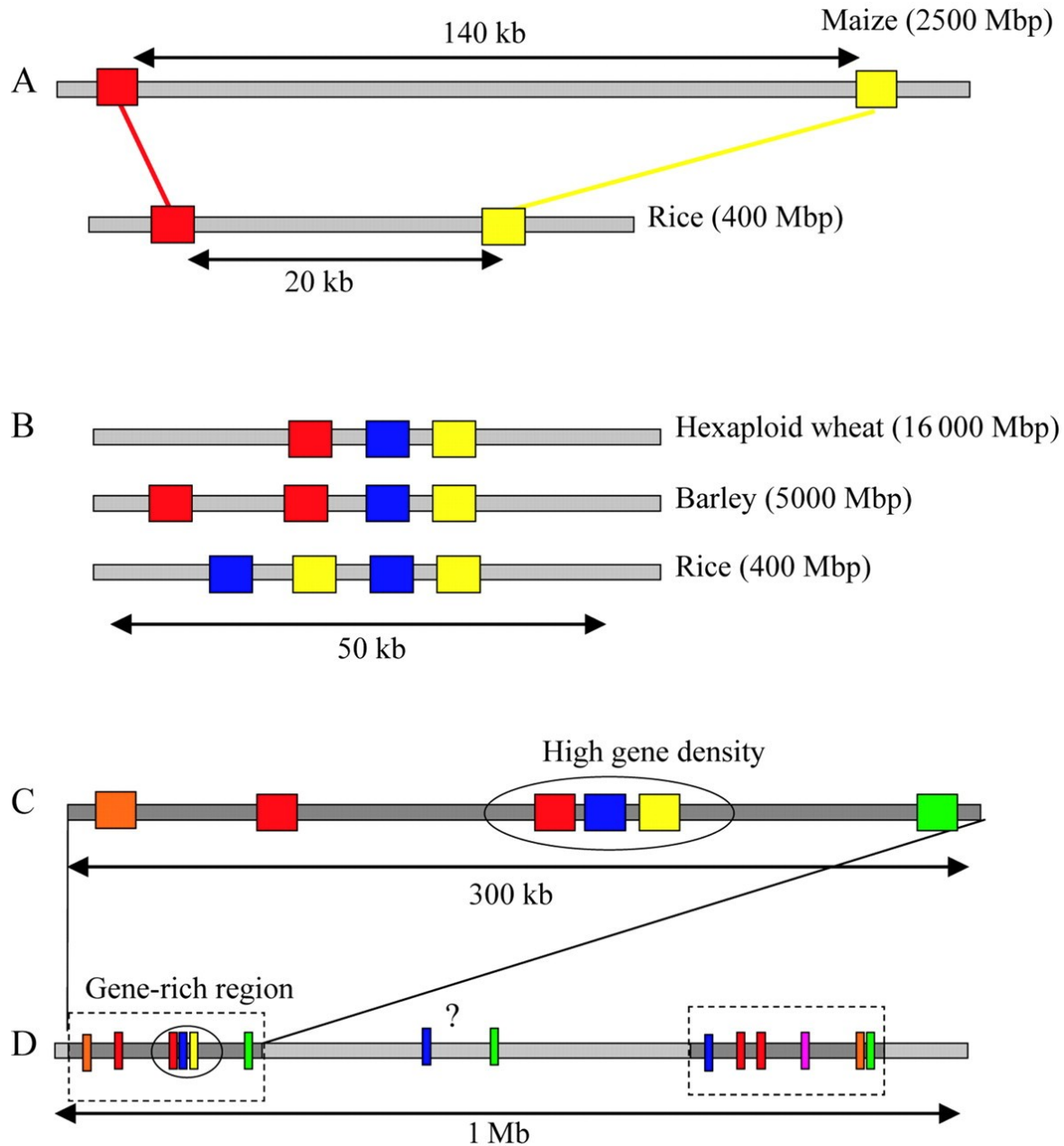
Osnova

- Identifikace genů *ab initio*
 - struktura genů a jejich vyhledávání
 - genomová kolinearita a genová homologie

Genomová kolinearita

- genomy příbuzných druhů se přes značné odlišnosti vyznačují podobnostmi v uspořádání i sekvencích, možnost využití při identifikaci genů u příbuzných organismů pomocí vyhledávání v databázích
- obecné schéma postupu při využívání genomové kolinearity (také „komparativní genomika“) při experimentální identifikaci genů příbuzných organismů:
 - mapování malých genomů s využitím nízkokopiových DNA markerů (např. RFLP)
 - využití těchto markerů k identifikaci orthologních genů (genů se stejnou nebo podobnou funkcí) příbuzného organismu
 - malý genom (např. rýže, 466 Mbp) může sloužit jako vodítko, kdy jsou identifikovány molekulární nízkokopiové markery (např. RFLP) ve vazbě s genem zájmu a tyto oblasti jsou pak použity jako sonda při vyhledávání v BAC knihovnách při identifikaci orthologních oblastí velkých genomů (např. ječmene nebo pšenice, 5000, resp. 16000 Mbp)

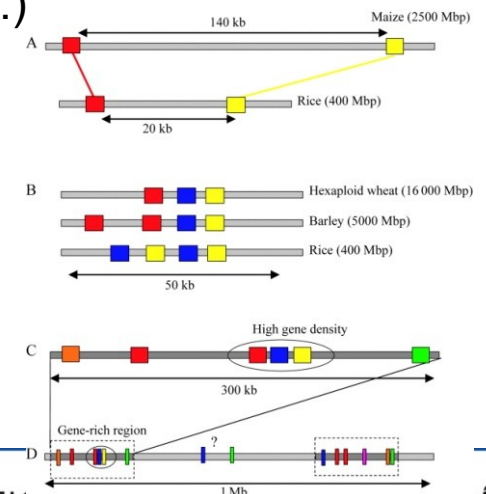
Genomová kolinearita



Feuillet and Keller, 2002

Genomová kolinearita

- zejména využitelné u trav (např. využití příbuznosti u ječmene, pšenice, rýže a kukuřice)
- malé genomové přestavby (dalece, duplikace, inverze a translokace menší než několik cM) jsou pak detekovány podrobnou sekvenční komparativní analýzou
- během evoluce dochází u příbuzných druhů k odchylkám především v nekódujících oblastech (invaze retrotranspozonů atd.)



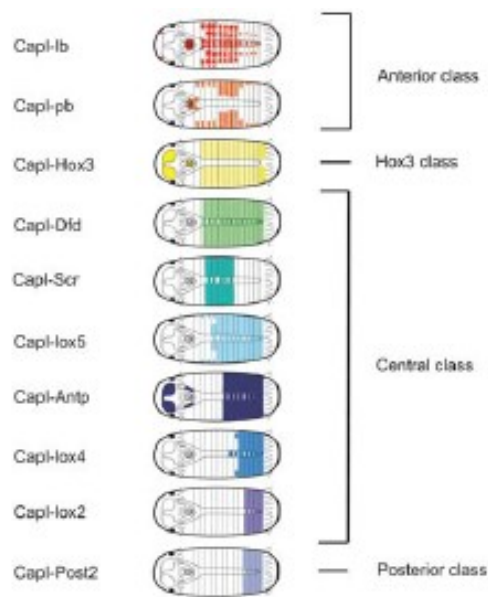
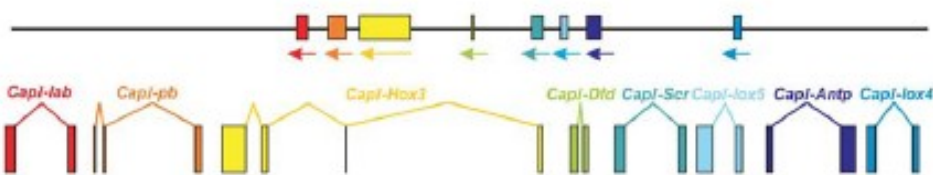
INVESTICE DO ROZVOJE VZDELÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Genomová kolinearita

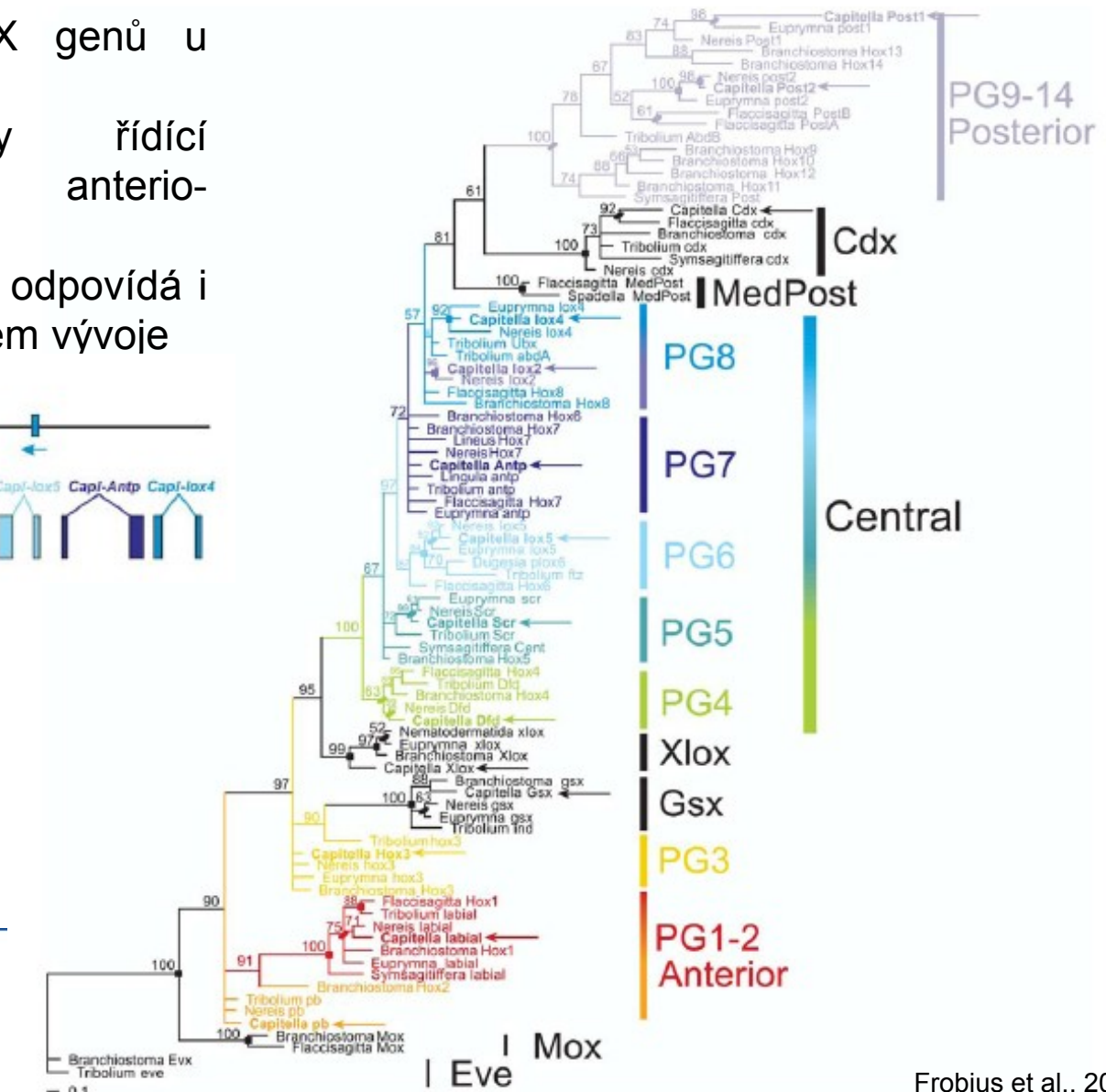
- Genomová kolinearita HOX genů u živočichů
 - Transkripční faktory řídící organizaci těla v antero-posteriorní ose
 - Pozice genů v genomu odpovídá i prostorové expresi během vývoje

Scaffold 70 499,360 bp



EVR

ŠKOLSTVÍ,
VÝCHOVY



Frobis et al., 2008

lí
ia
n

Osnova

- Experimentální identifikace genů
 - příprava genově obohacených knihoven pomocí technologie metylačního filtrování

Metylační filtrování

- příprava genově obohacených knihoven pomocí technologie metylačního filtrování
- geny jsou (většinou!) hypometylované, kdežto nekódující oblasti jsou metylované
- využití bakteriálního RM systému, který rozpoznává metylovanou DNA pomocí rest. enzymů McrA a McrBC
 - McrBC rozpoznává v DNA metylovaný cytozin, který předchází purin (G nebo A)
 - pro štěpení je nutná vzdálenost těchto míst z 40-2000 bp

Metylační filtrování

- příprava genově obohacených knihoven pomocí technologie metylačního filtrování
- schéma postupu při přípravě BAC genomových knihoven pomocí metylačního filtrování:
 - příprava genomové DNA bez příměsí organelární DNA (chloroplasty a mitochondrie)
 - fragmentace DNA (1-4 kbp) a ligace adaptorů
 - příprava BAC knihovny v *mcrBC+* kmeni *E. coli*
 - selekce pozitivních klonů
- omezené využití: obohacení o kódující DNA o pouze cca 5-10 %

Osnova

- Experimentální identifikace genů
 - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
 - EST knihovny

EST knihovny

- příprava EST knihoven

- izolace mRNA
- RT PCR
- ligace linkerů a syntéza druhého řetězce cDNA
- klonování do vhodného bakteriálního vektoru
- transformace do bakterií a izolace DNA (amplifikace DNA)
- sekvenace s použitím specifických primerů pro použitý plasmid
- uložení výsledků sekvenace do veřejné databáze

NCBI Nucleotide

Search Nucleotide for

Display [default] Show [20] Send to [File] Get Subsequence Features

1: NC_002377.1 Agrobacterium tum. [gi:10955016]

LOCUS NC_002377 2490 bp DNA linear BCT 28-DEC-2003

DEFINITION Agrobacterium tumefaciens cosmid Ti, complete sequence.

ACCESSION NC_002377 REGION 14694..14813

VERSION NC_002377.1 GI:10955016

FEATURES

1. Agrobacterium tumefaciens (Rhizobium radiobacter)

ORGANISM Agrobacterium tumefaciens

Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Rhizobiales; Rhizobium/Agrobacterium group; Agrobacterium.

REFERENCE 1 (bases 1 to 2490)

AUTHORS Winans, S.C., Zhu, J., Oper, P.M., Schrammeyer, B., Hooyskaas, P.J. and Parrand, E.K.

TITLE Octopine-type Ti plasmid sequence

JOURNAL Unpublished

REFERENCE 2 (bases 1 to 2490)

AUTHORS Zhu, J., Oper, P.M., Schrammeyer, B., Hooyskaas, P.J., Parrand, E.K. and Winans, S.C.

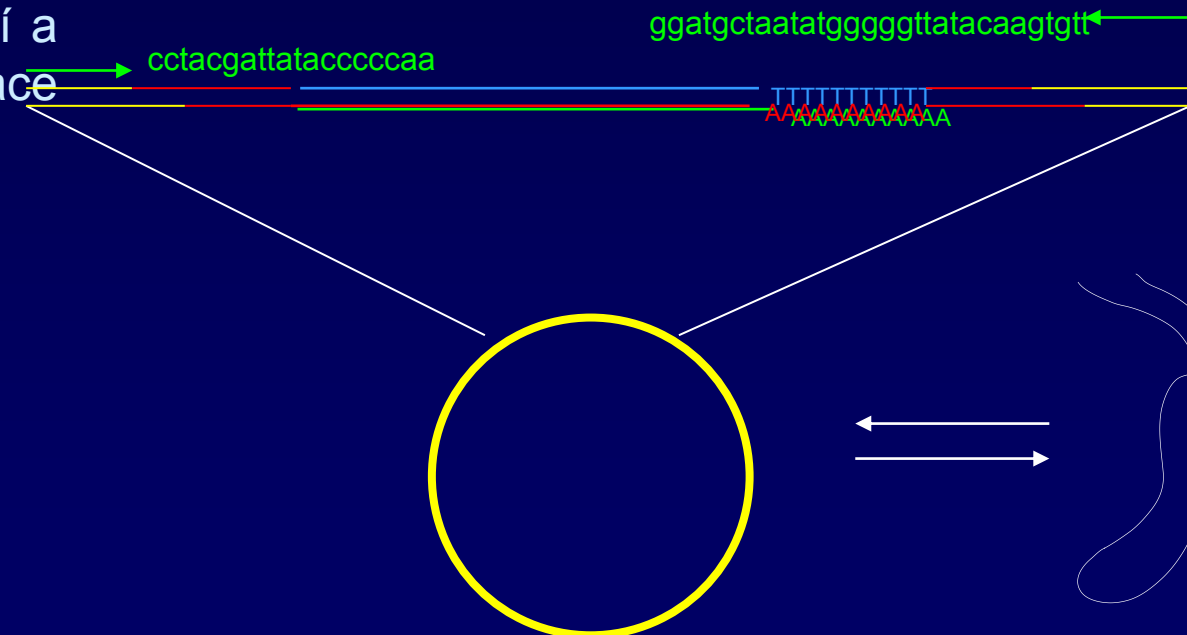
TITLE Direct Submission

JOURNAL Submitted (07-MAR-2003) Microbiology, Cornell University, Wing Hall, Ithaca, NY 14853, USA

COMMENT PROVISIONAL REFSEQ: This record has not yet been subject to final NCBI review. The reference sequence was derived from [SRR1182](#).

FEATURES

source	1..2490	/organism="Agrobacterium tumefaciens"
		/mol_type="genomic DNA"
		/db_xref="taxon:216"
		/plasmid="Ti"
		/note="extrachromosomal"
gene	1..2490	octopine-type"
		/gene="virA"
		/db_xref="GeneID:1224316"
CDS	1..2490	/gene="virA"
		/note="two-component regulator of vir regulon; VirA is a transmembrane histidine kinase"
		/codon_start=1
		/transl_table=11
		/product="VirA"
		/protein_id="YP_002377.1"
		/db_xref="GI:10955016"



Osnova

- Postupy „přímé“ a reverzní genetiky
 - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
 - struktura genů a jejich vyhledávání
 - genomová kolinearita a genová homologie
- Experimentální identifikace genů
 - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
 - EST knihovny
 - přímá a reverzní genetika

Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky
 - Změna fenotypu po mutagenezi
 - **Genetika přímá**
 - Identifikace inzerčního mutanta a analýza jeho fenotypu
 - **Genetika reverzní**
 - Analýza exprese daného genu a jeho časoprostorové specifiity

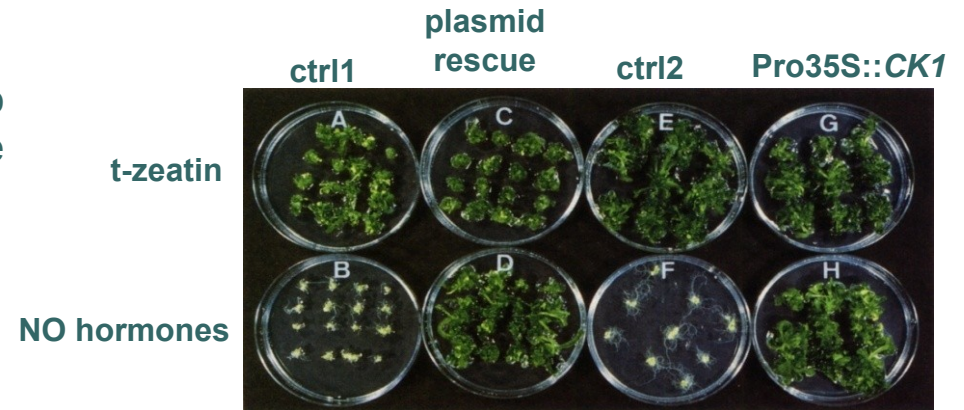
Přímá a reverzní genetika - shrnutí

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky
 - Změna fenotypu po mutagenezi
 - **Genetika přímá**

Cloning of CKI1

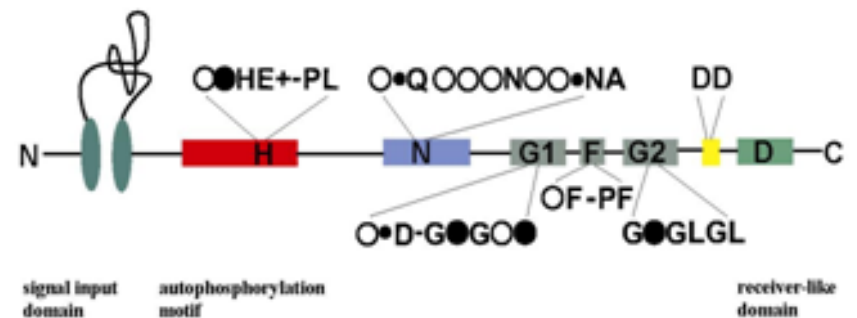
- CKI1 was identified via activation mutagenesis in *Arabidopsis*

- Overexpression of *CKI1* leads to CK-like response in the hypocotyl explants



Kakimoto, *Science* (1996)

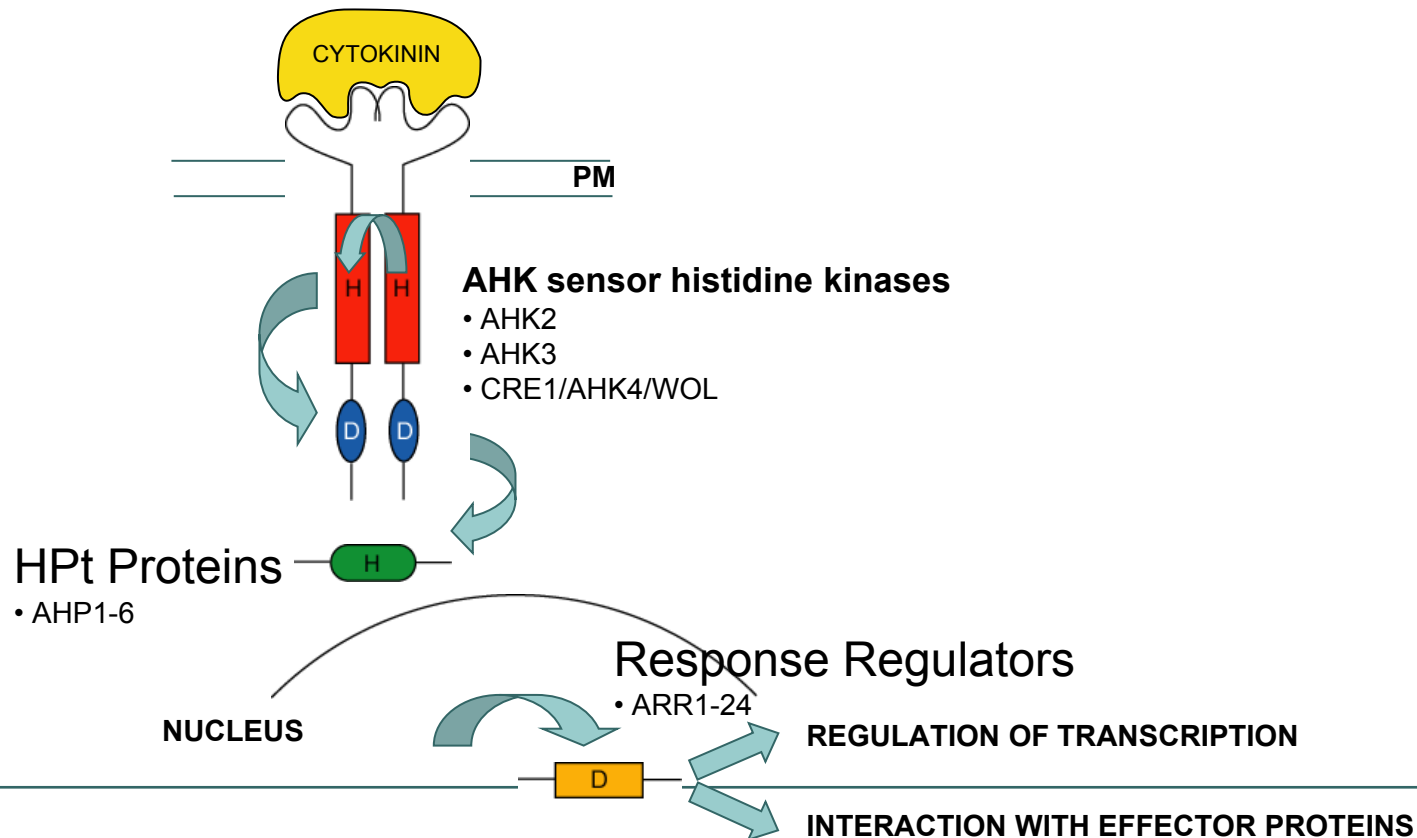
- CKI1* encodes a protein with similarity to bacterial histidine kinases



- O non-polar amino acids
- polar amino acids (a.a.)
- + basic a.a.
- acidic or amidic a.a.
- positions with less than 50% conservation

Signal Transduction via MSP

Recent Model of the CK Signaling via Multistep Phosphorelay (MSP) Pathway





Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky
 - Změna fenotypu po mutagenezi
 - **Genetika přímá**
 - Identifikace inzerčního mutanta a analýza jeho fenotypu
 - **Genetika reverzní**
-

Identification of insertional *cki1* mutant allele



aattcaagtcgctCACTACAAGA " **En-1** TCTTGTAGTGCgtggagact

A. aat tca agt **cg t gga gac tac** act tgg tac act caa acc gtg gat cag tta act ggt
N S S **R G D Y** T W Y T Q T V D Q L T G

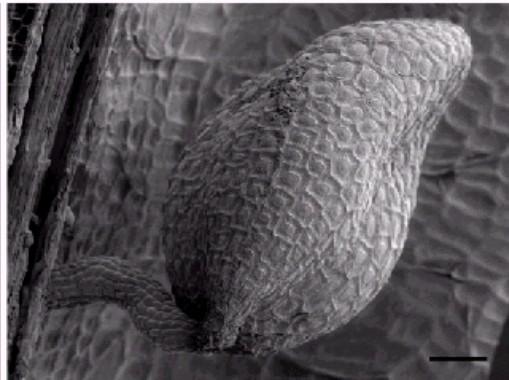
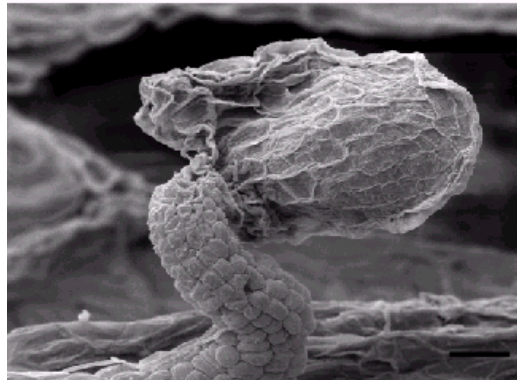
CKI1 Regulates Female Gametophyte Development

- CKI1 is necessary for proper megagametogenesis in *Arabidopsis*

CKI1/cki1-i



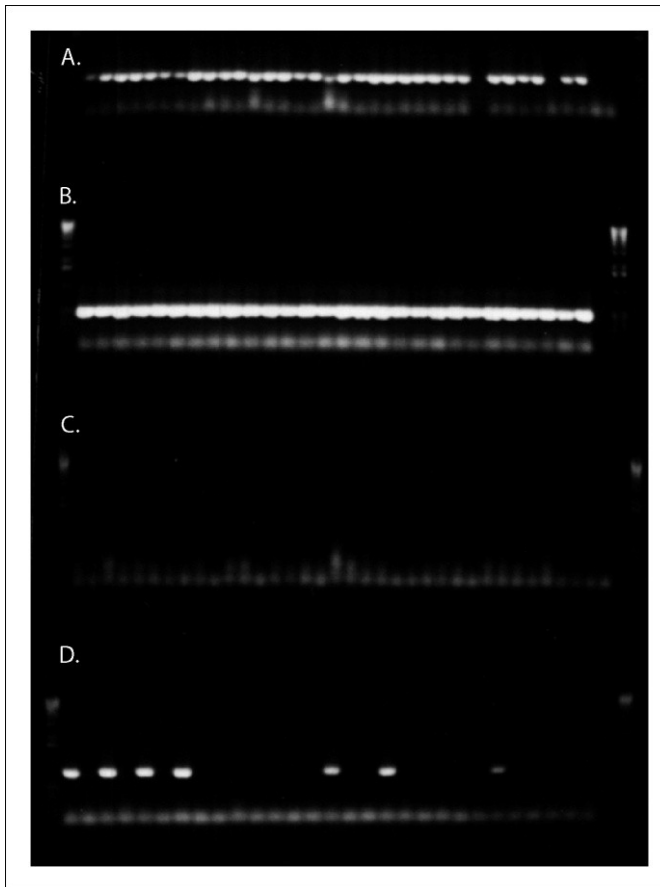
CKI1/CKI1



Hejátko et al., *Mol Genet Genomics* (2003)

CKI1 and megagametogenesis

- *cki1-i* is not transmitted through the female gametophyte



A. wt x *CKI1/cki1-i*



CKI1 specific primers (PCR positive control)

B. *CKI1/cki1-i* x wt

C. wt x *CKI1/cki1-i*

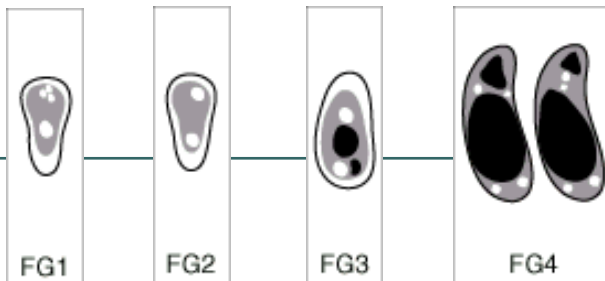
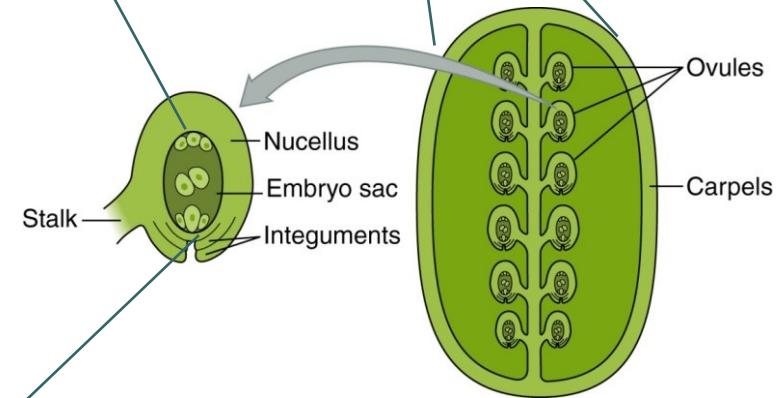
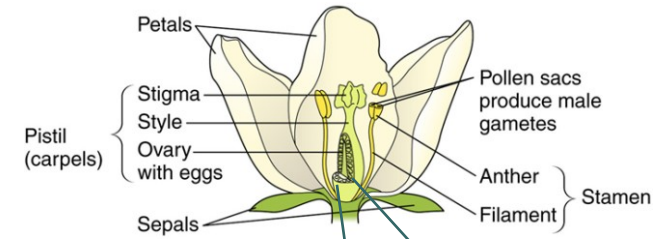
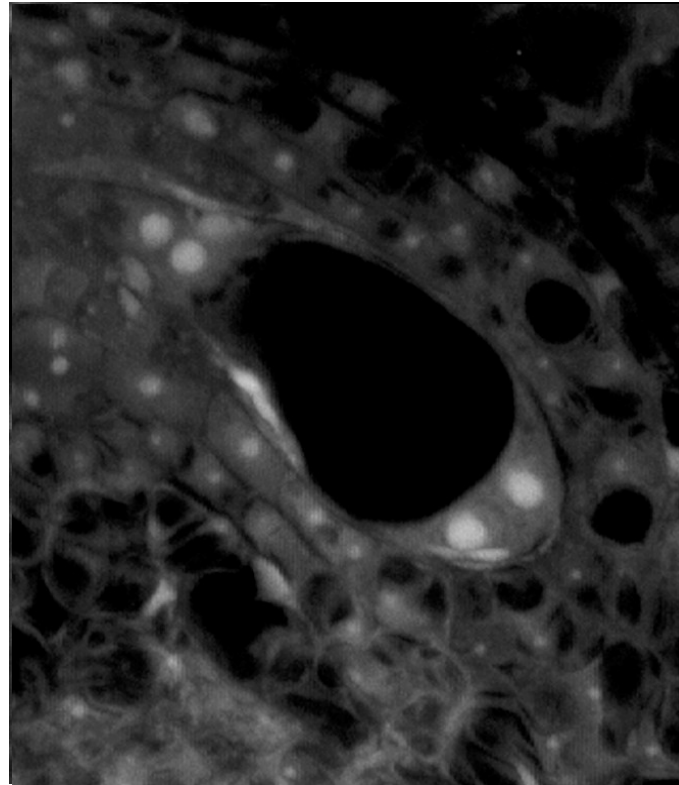


cki1-i specific primers

D. *CKI1/cki1-i* x wt

CKI1 and megagametogenesis

FG 4



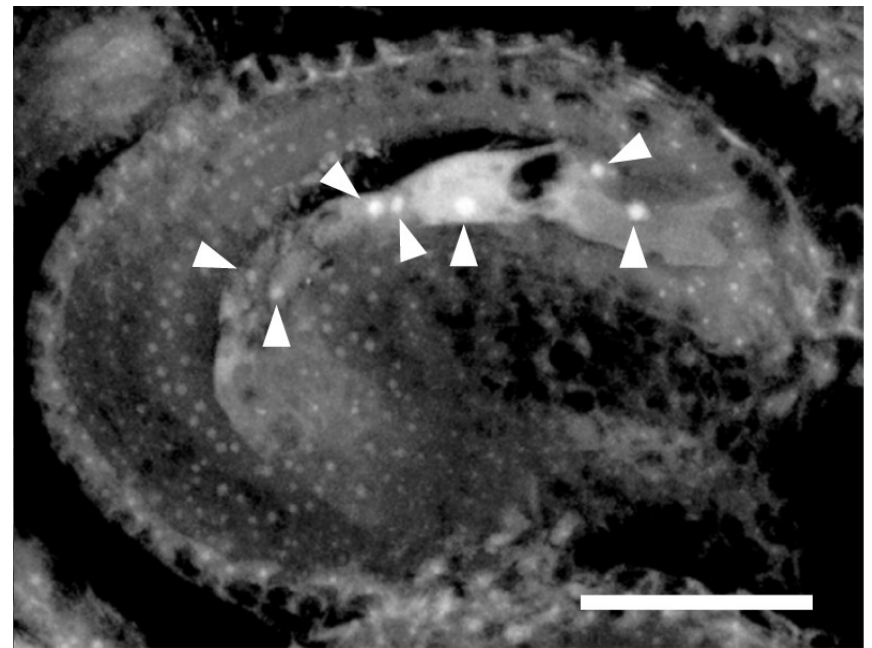
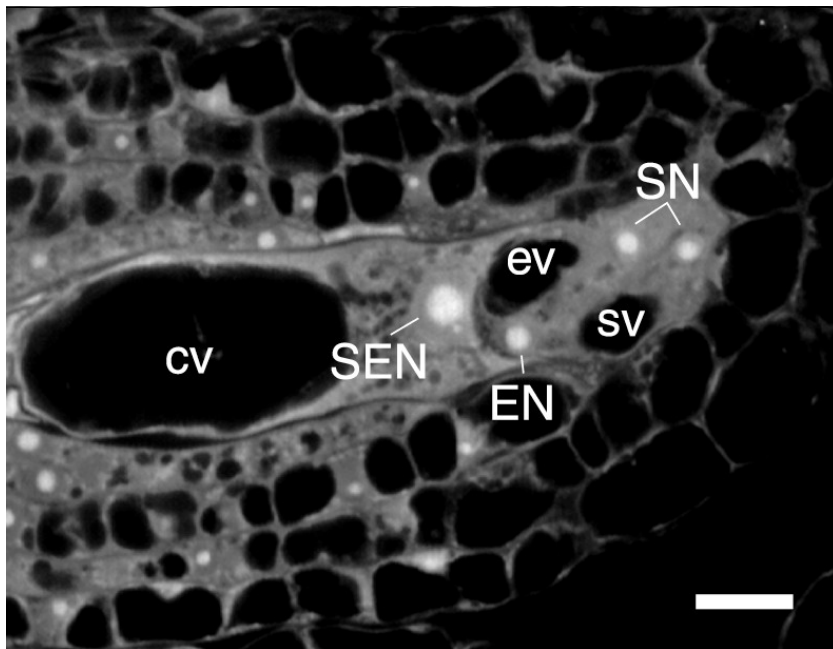
CKI1 and megagametogenesis

CKI1

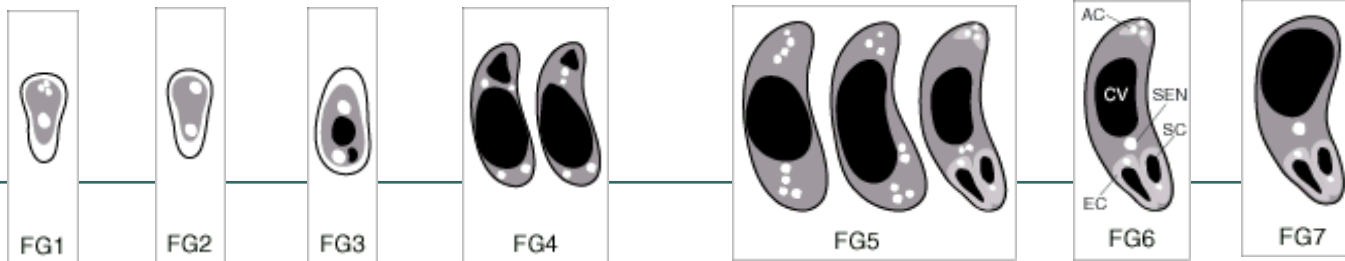
FG4 to FG5

cki1-i

28 HAE



Hejátko et al., *Mol Genet Genomics* (2003)

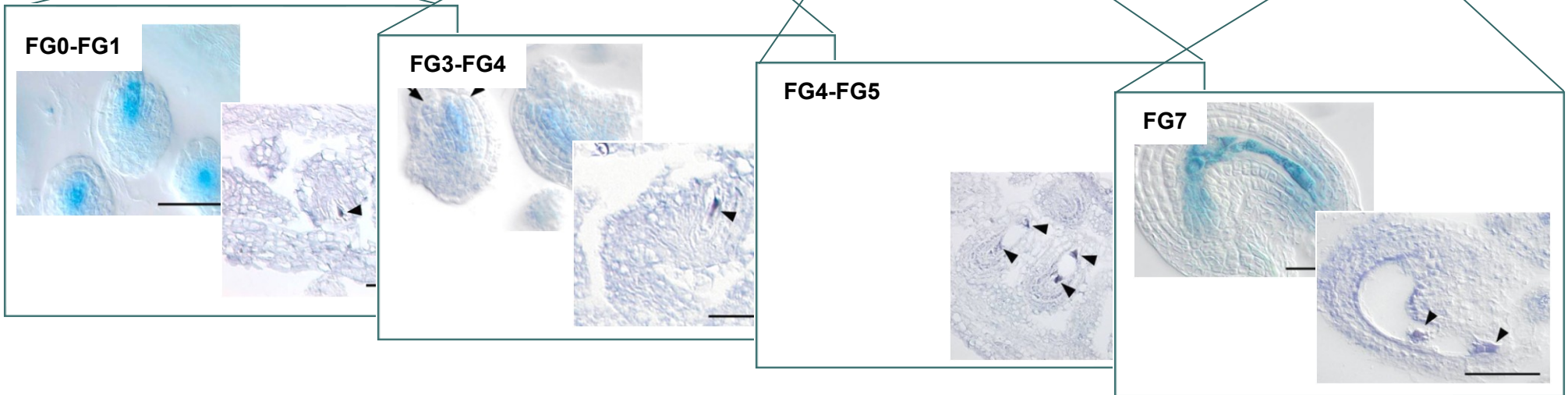
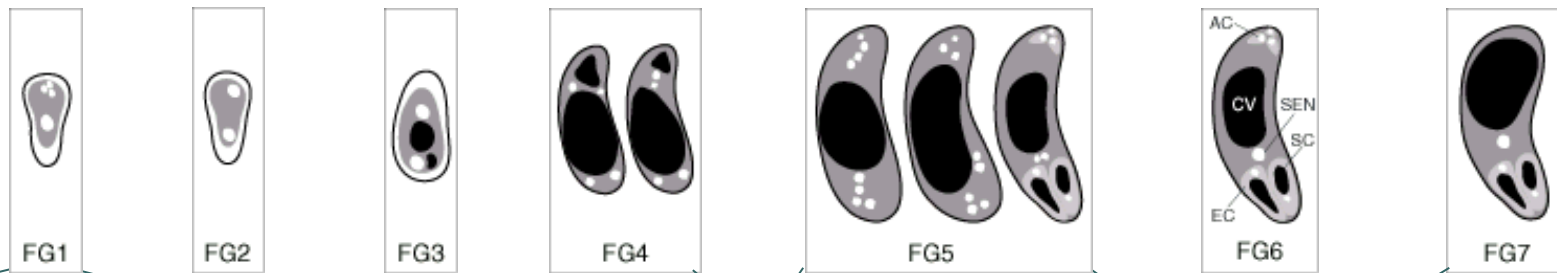




Přímá a reverzní genetika

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky
 - Změna fenotypu po mutagenezi
 - **Genetika přímá**
 - Identifikace inzerčního mutanta a analýza jeho fenotypu
 - **Genetika reverzní**
 - Analýza exprese daného genu a jeho časoprostorové specifity
-

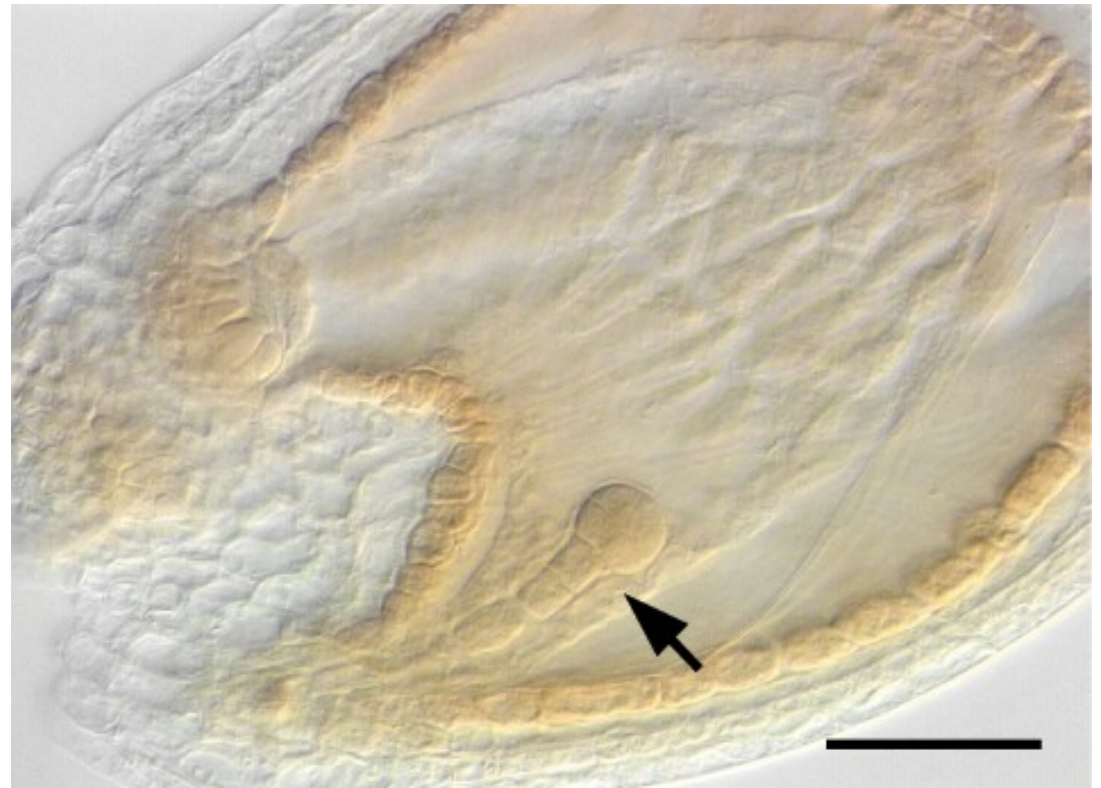
CKI1 is Expressed Throughout Megagametogenesis

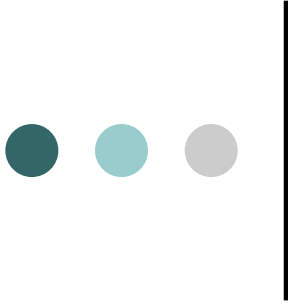


Paternal *CKI1* is Expressed in the *Arabidopsis* Sporophyte Early after Fertilization

wt x Pro*CKI1*:*GUS*

22 HAP
(hours
after
pollination)





Přímá a reverzní genetika - shrnutí

- Principy experimentální identifikace genů prostřednictvím přímé a reverzní genetiky
 - Změna fenotypu po mutagenezi
 - **Genetika přímá**
 - Identifikace inzerčního mutanta a analýza jeho fenotypu
 - **Genetika reverzní**
 - Analýza exprese daného genu a jeho časoprostorové specifity
-

Shrnutí

- Postupy „přímé“ a reverzní genetiky
 - rozdíly v myšlenkových přístupech k identifikaci genů a jejich funkcí
- Identifikace genů *ab initio*
 - struktura genů a jejich vyhledávání
 - genomová kolinearita a genová homologie
- Experimentální identifikace genů
 - příprava genově obohacených knihoven pomocí technologie metylačního filtrování
 - EST knihovny
 - přímá a reverzní genetika

Diskuse



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky