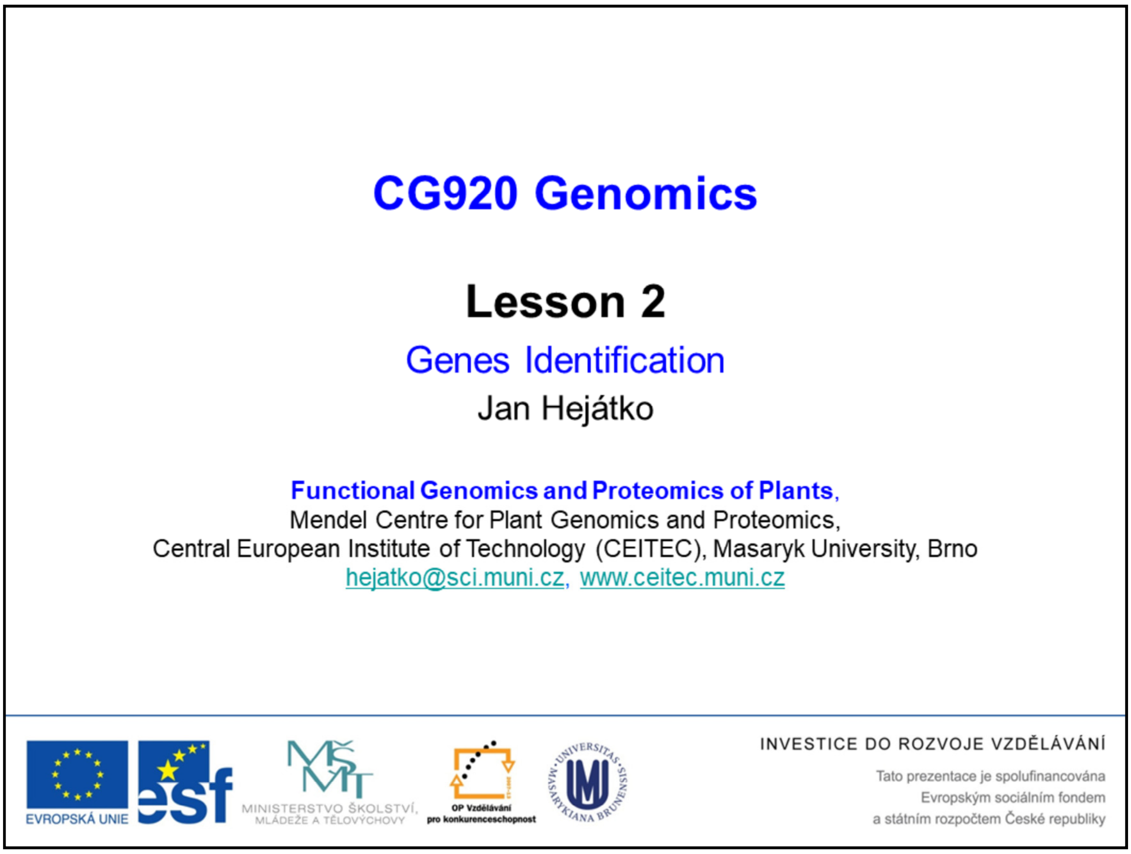


CG920 Genomics

Lesson 2

Genes Identification

Jan Hejátko

Functional Genomics and Proteomics of Plants,
Mendel Centre for Plant Genomics and Proteomics,
Central European Institute of Technology (CEITEC), Masaryk University, Brno
hejatko@sci.muni.cz, www.ceitec.muni.cz

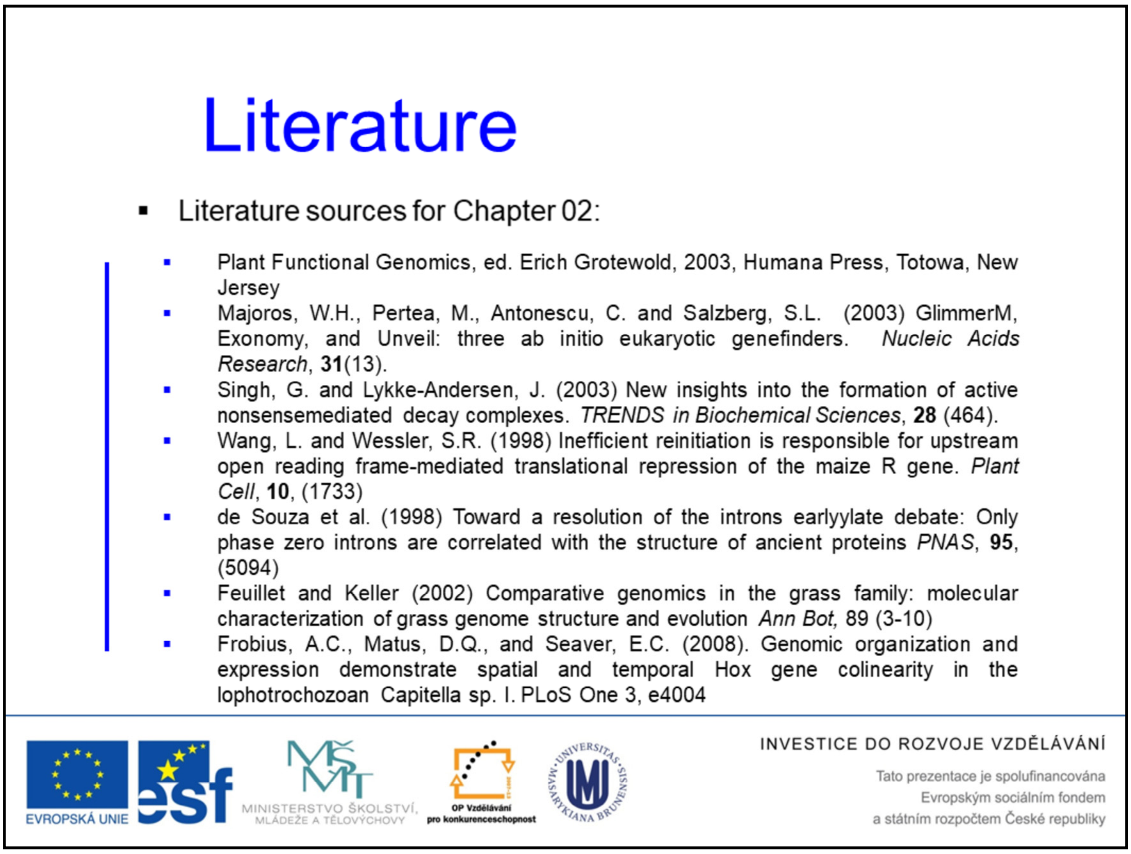
EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



Literature

- Literature sources for Chapter 02:

- 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 nonsensemediated 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 earlyylate 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)
- Frobius, 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

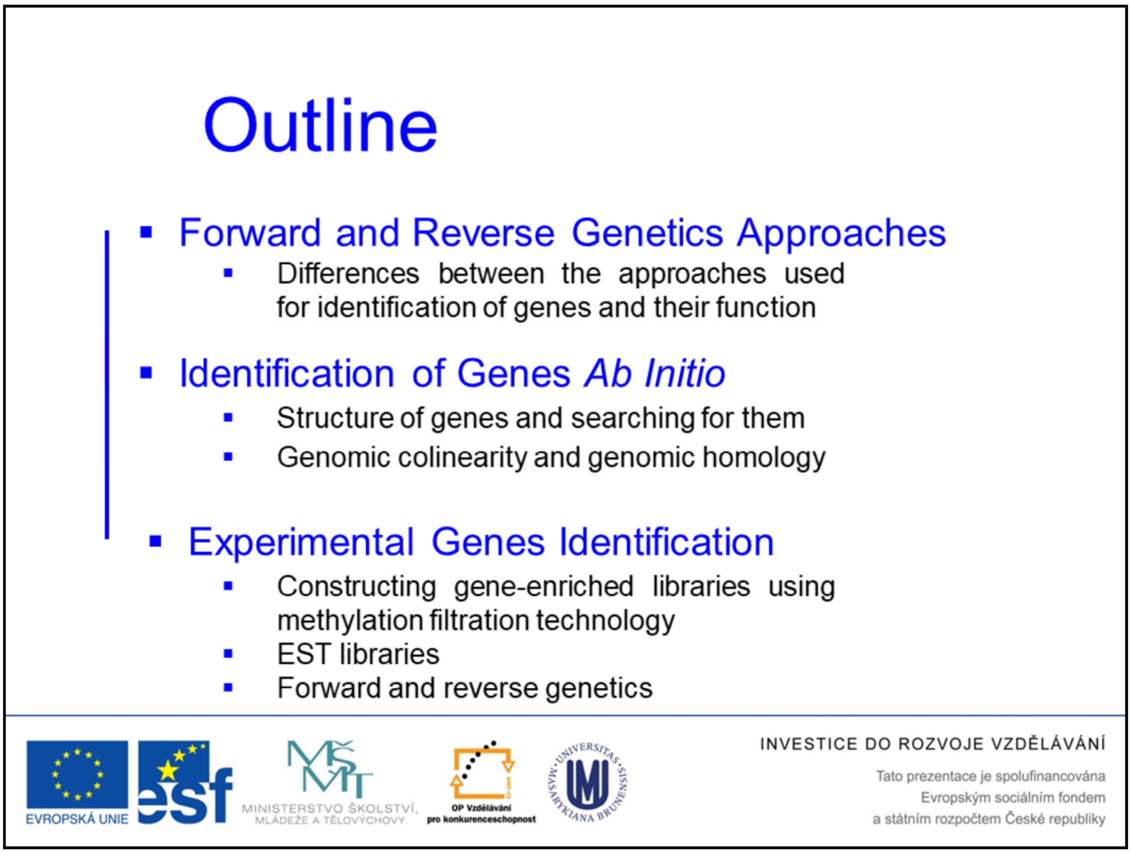
EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

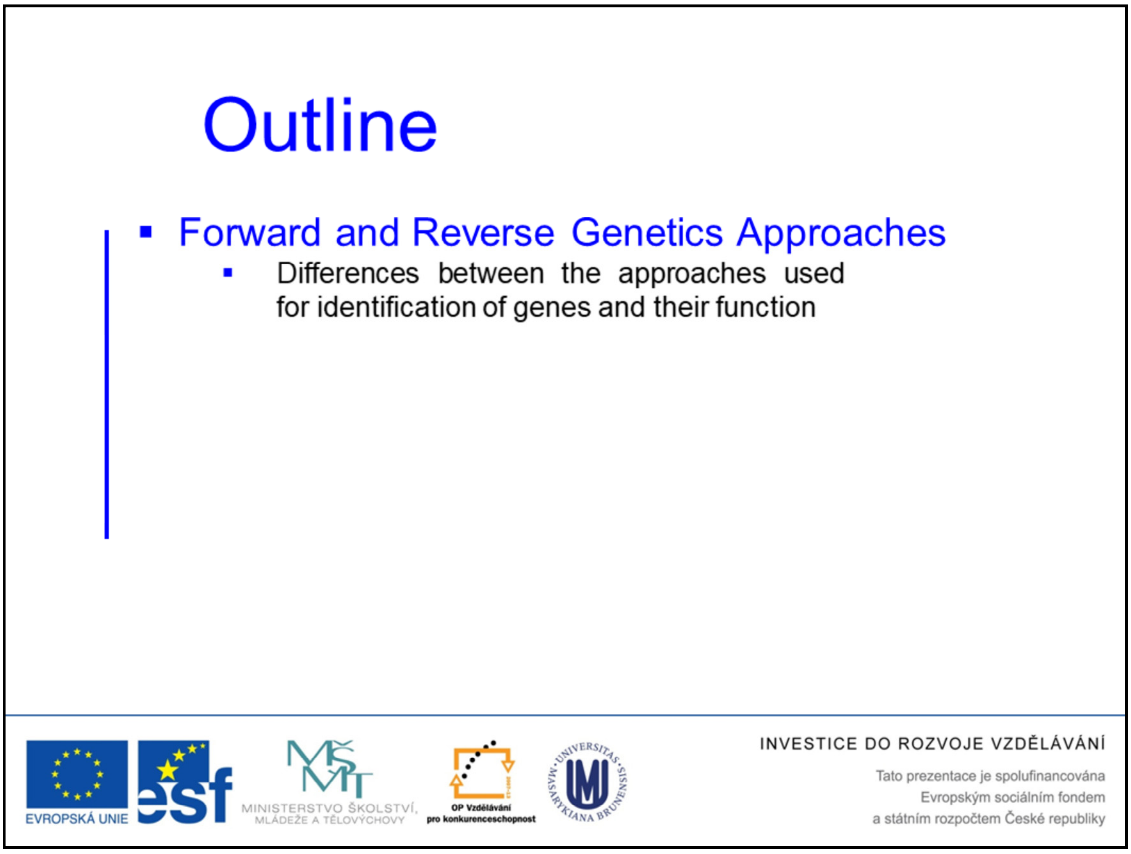


Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them
  - Genomic colinearity and genomic homology
- Experimental Genes Identification
  - Constructing gene-enriched libraries using methylation filtration technology
  - EST libraries
  - Forward and reverse genetics

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

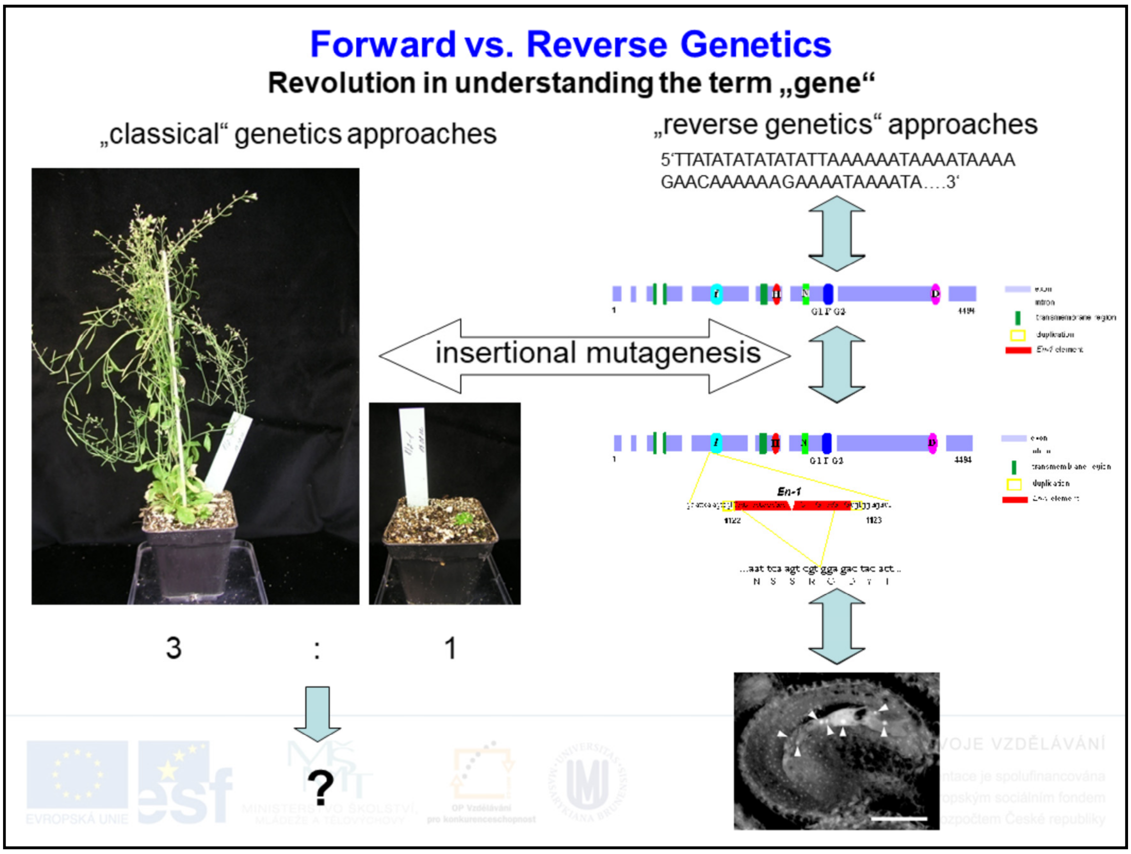


Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



Forward vs. Reverse Genetics

Revolution in understanding the term „gene“

„classical“ genetics approaches

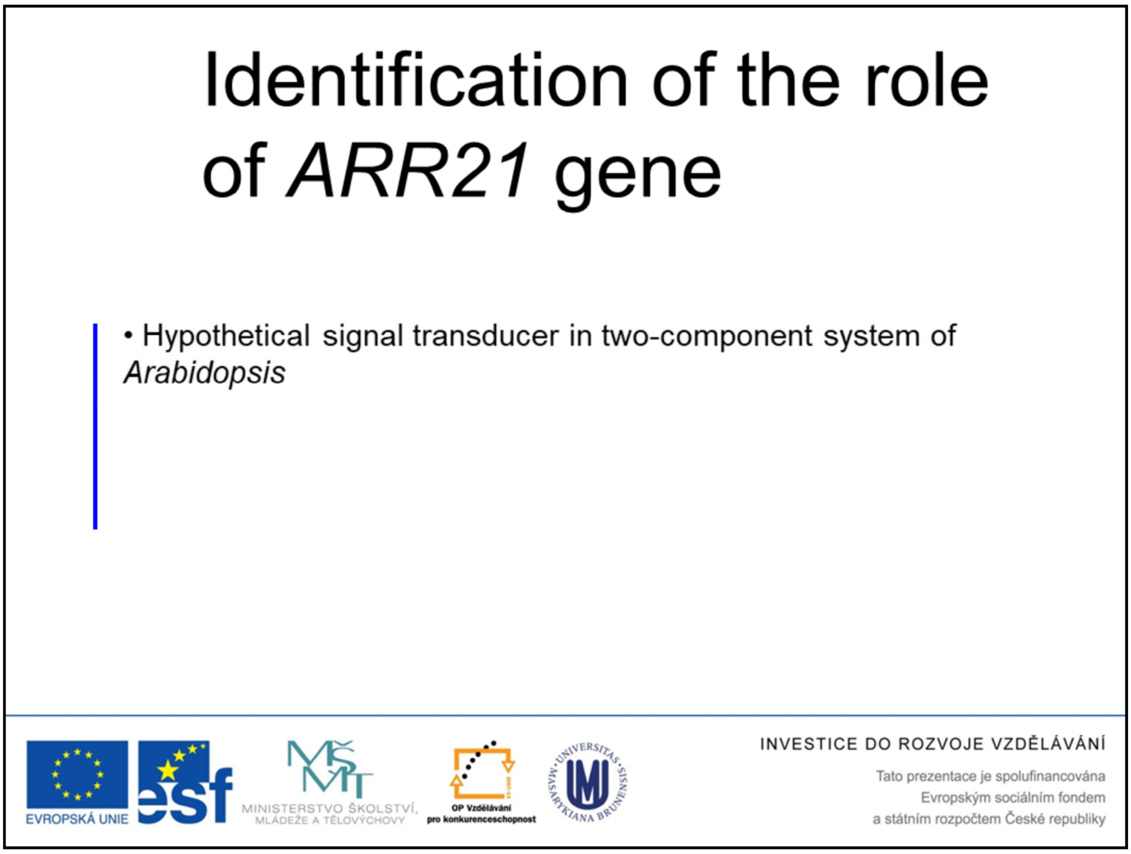
„reverse genetics“ approaches

5‘TTATATATATATATTAAAAAATAAAATAAAA GAACAAAAAAGAAAATAAAATA….3‘

insertional mutagenesis

3 : 1

?

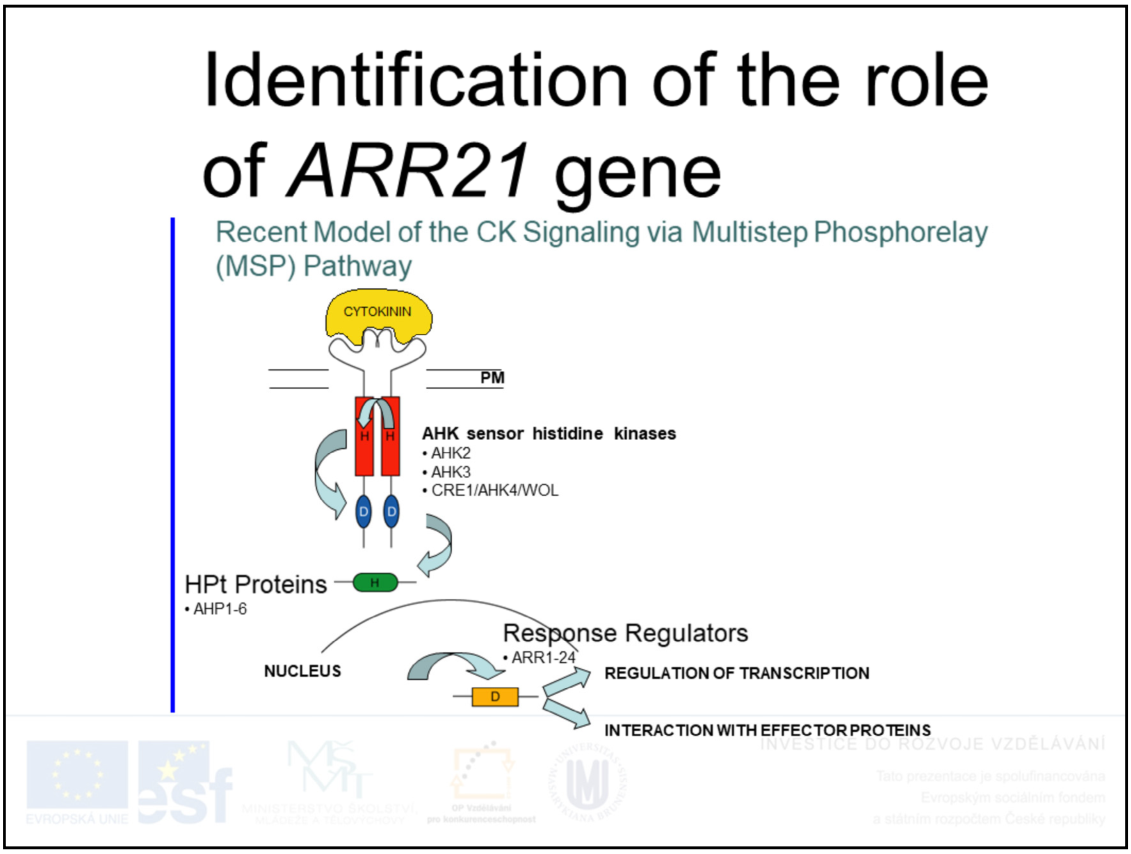


Identification of the role of *ARR21* gene

- Hypothetical signal transducer in two-component system of *Arabidopsis*

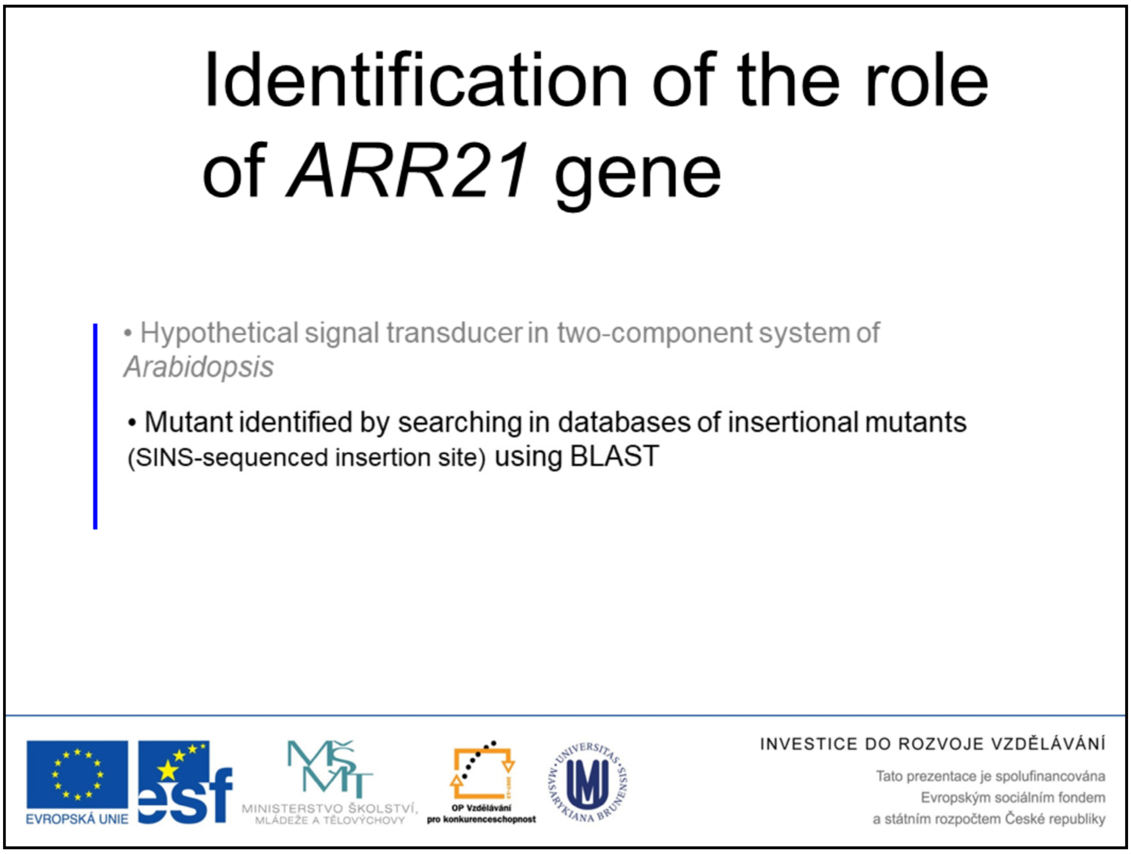
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



Identification of the role of *ARR21* gene

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

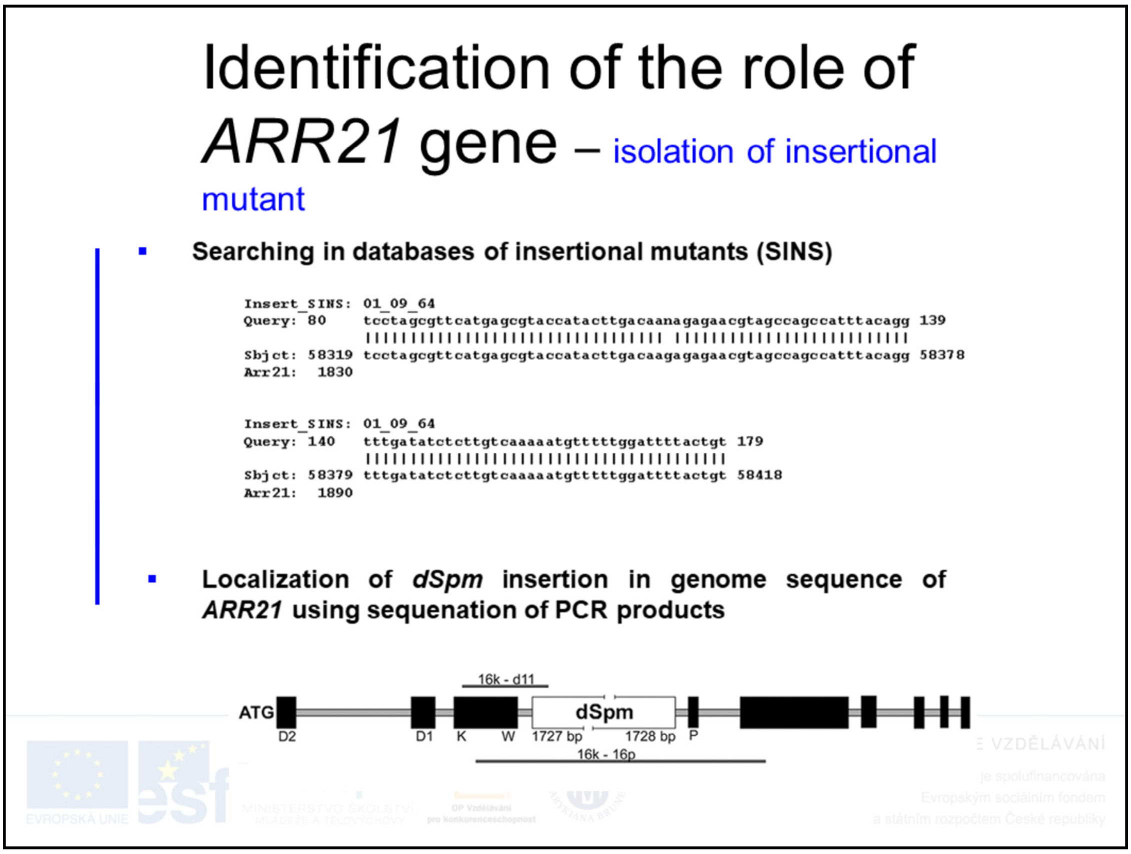


Identification of the role of *ARR21* gene

- Hypothetical signal transducer in two-component system of *Arabidopsis*
- Mutant identified by searching in databases of insertional mutants (SINS-sequenced insertion site) using BLAST

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

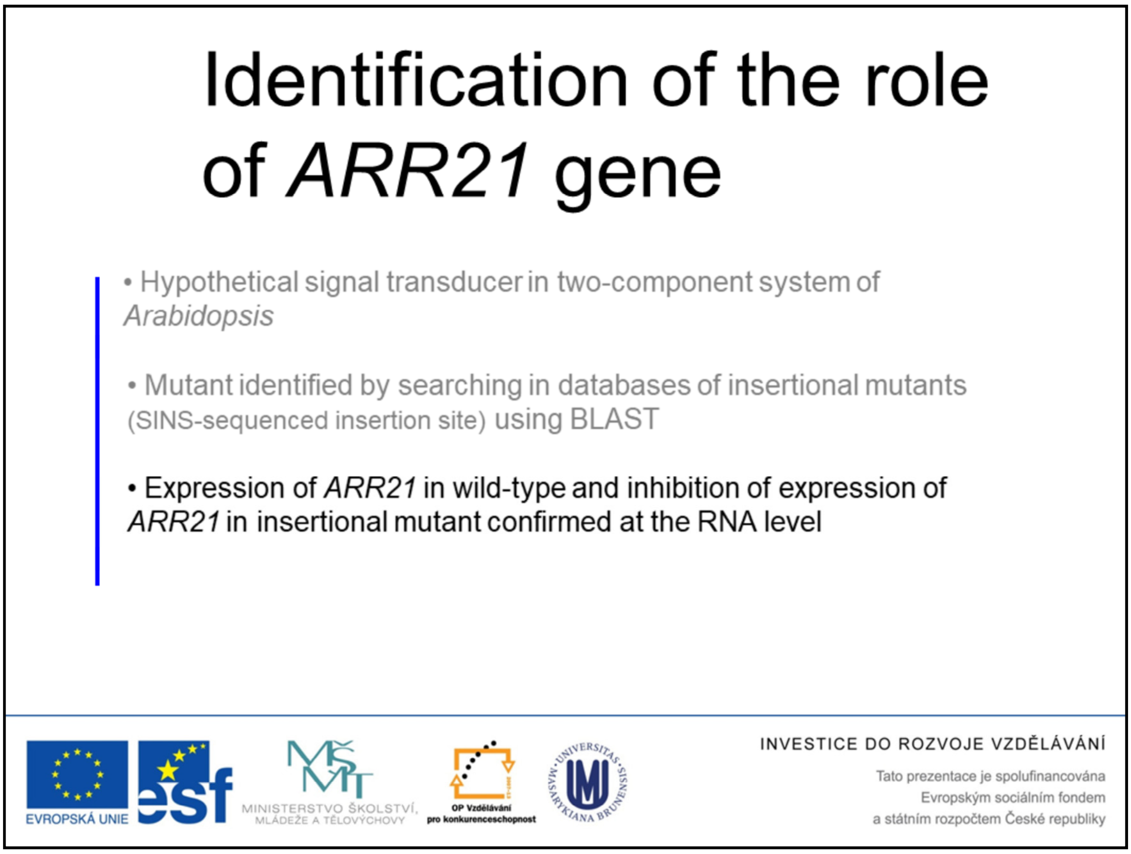


Identification of the role of *ARR21* gene – isolation of insertional mutant

- Searching in databases of insertional mutants (SINS)

```
Insert_SINS: 01_09_64
Query: 80    tcctagcgttcatgagcgtaccatacttgacaanagagaacgtagccagccatttacagg 139
             ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct: 58319 tcctagcgttcatgagcgtaccatacttgacaagagagaacgtagccagccatttacagg 58378
Arr21:  1830


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

- Localization of *dSpm* insertion in genome sequence of *ARR21* using sequenation of PCR products



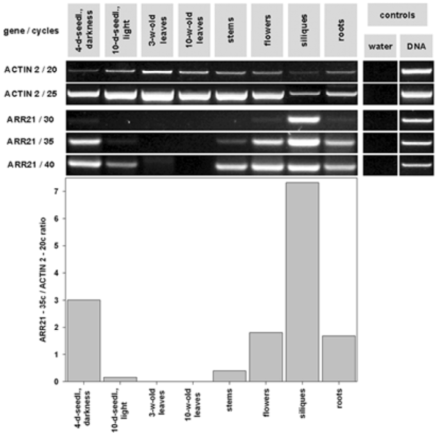
Identification of the role of *ARR21* gene

- Hypothetical signal transducer in two-component system of *Arabidopsis*
- Mutant identified by searching in databases of insertional mutants (SINS-sequenced insertion site) using BLAST
- Expression of *ARR21* in wild-type and inhibition of expression of *ARR21* in insertional mutant confirmed at the RNA level

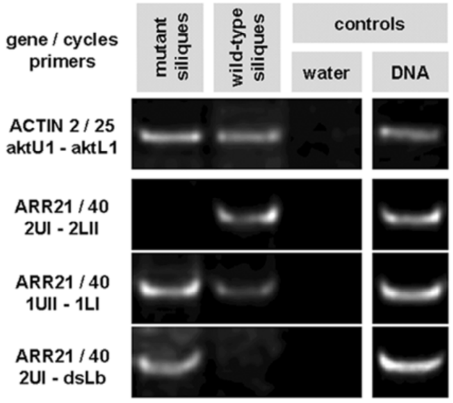
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Identification of the role of *ARR21* gene – analysis of expression

wild type expression

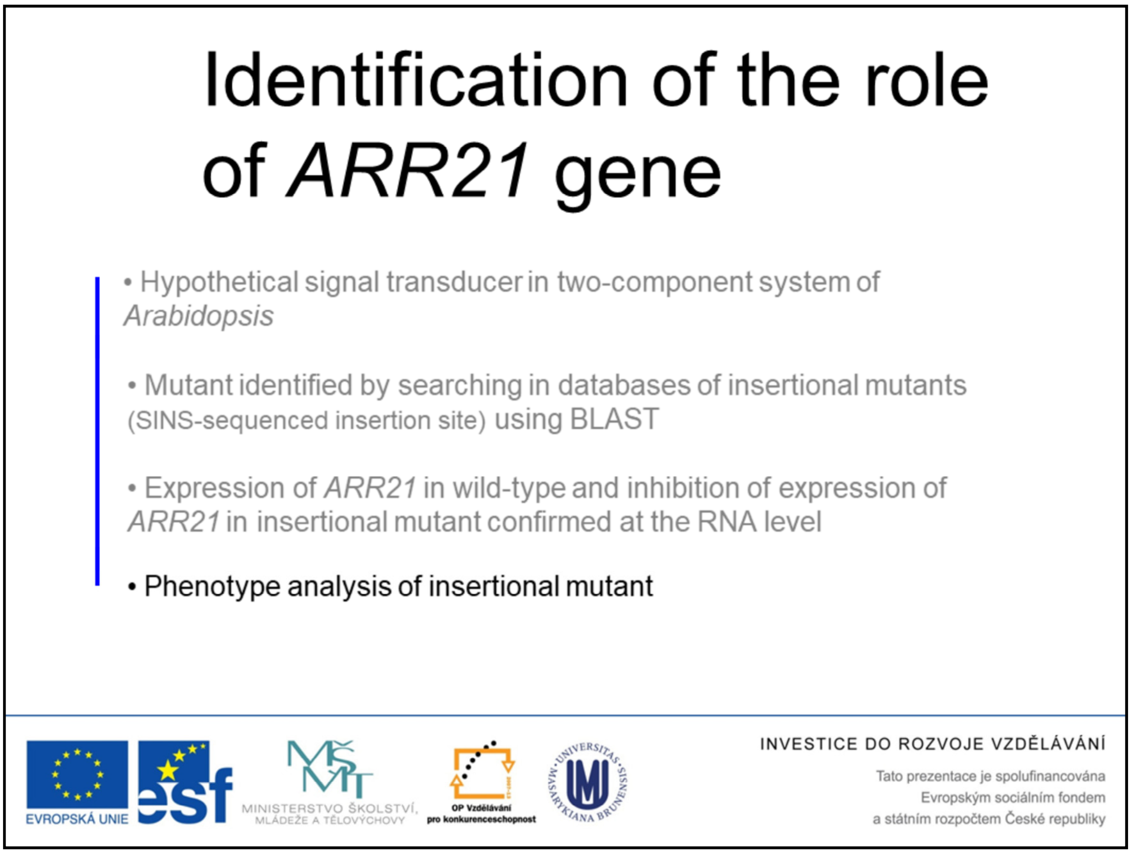


insertional mutant vs wild type



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

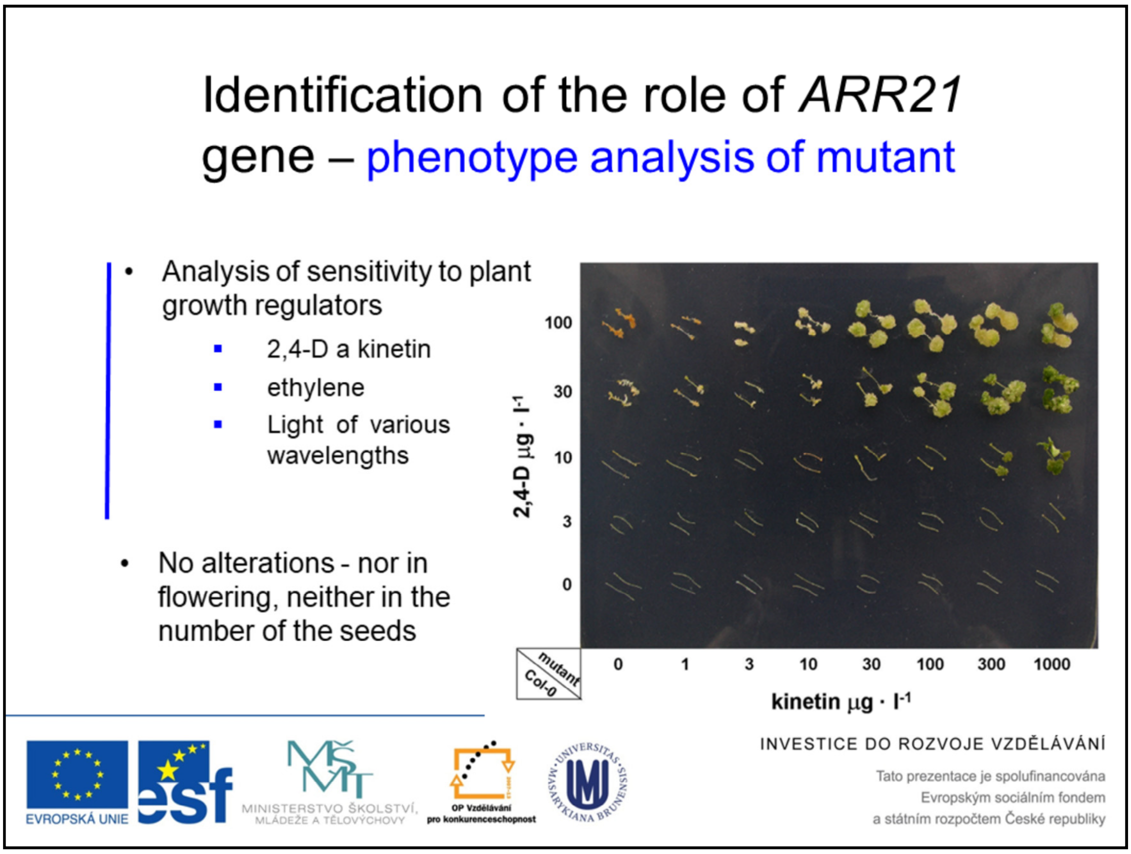
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Identification of the role of *ARR21* gene

- Hypothetical signal transducer in two-component system of *Arabidopsis*
- Mutant identified by searching in databases of insertional mutants (SINS-sequenced insertion site) using BLAST
- Expression of *ARR21* in wild-type and inhibition of expression of *ARR21* in insertional mutant confirmed at the RNA level
- Phenotype analysis of insertional mutant

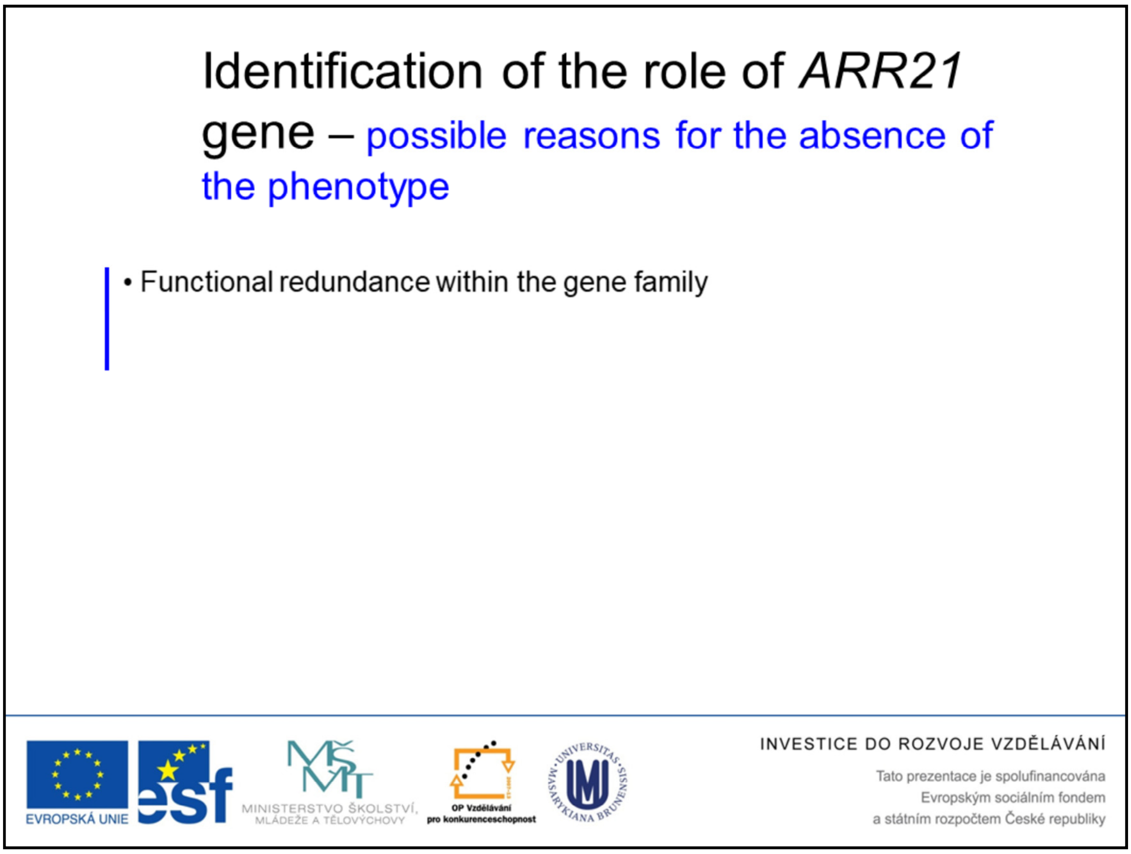
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Identification of the role of *ARR21* gene – phenotype analysis of mutant

- Analysis of sensitivity to plant growth regulators
  - 2,4-D a kinetin
  - ethylene
  - Light of various wavelengths
- No alterations - nor in flowering, neither in the number of the seeds

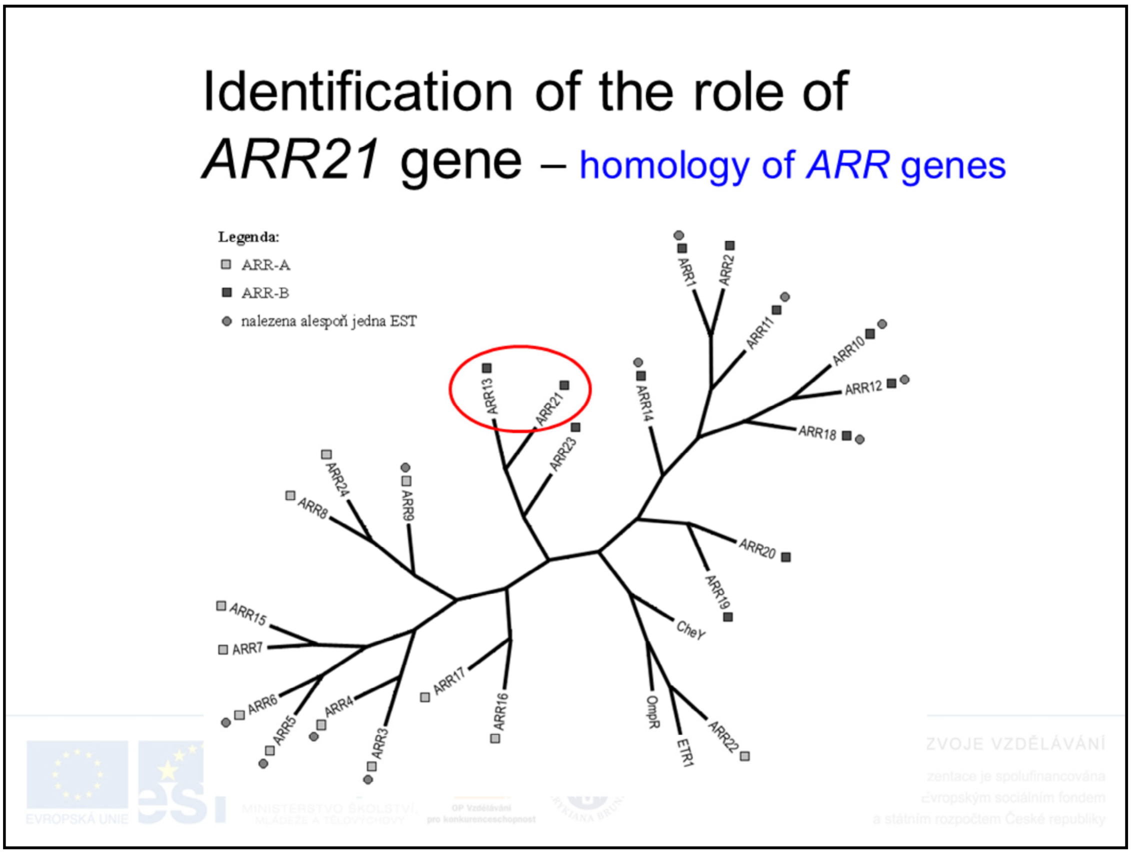
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Identification of the role of *ARR21* gene – possible reasons for the absence of the phenotype

- Functional redundance within the gene family

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



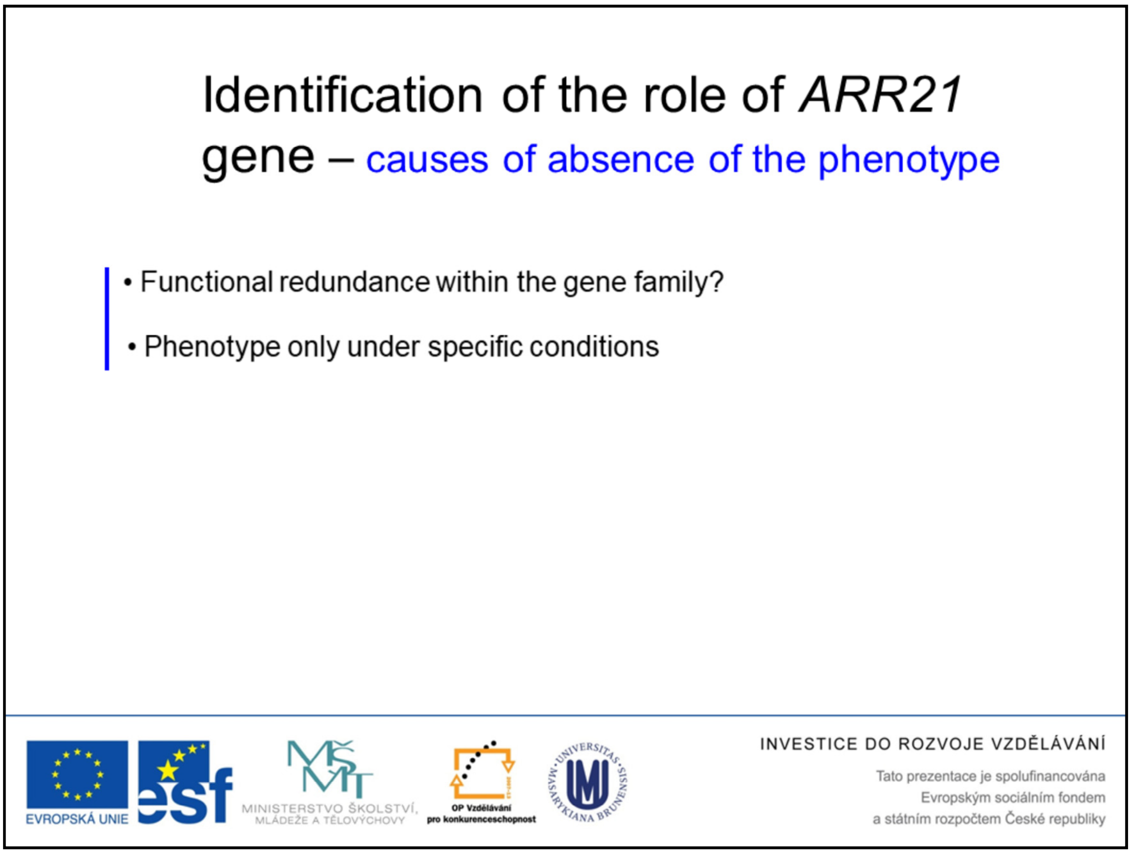
Identification of the role of *ARR21* gene – homology of *ARR* genes

Legenda:

□ ARR-A

■ ARR-B

● nalezena alespoň jedna EST

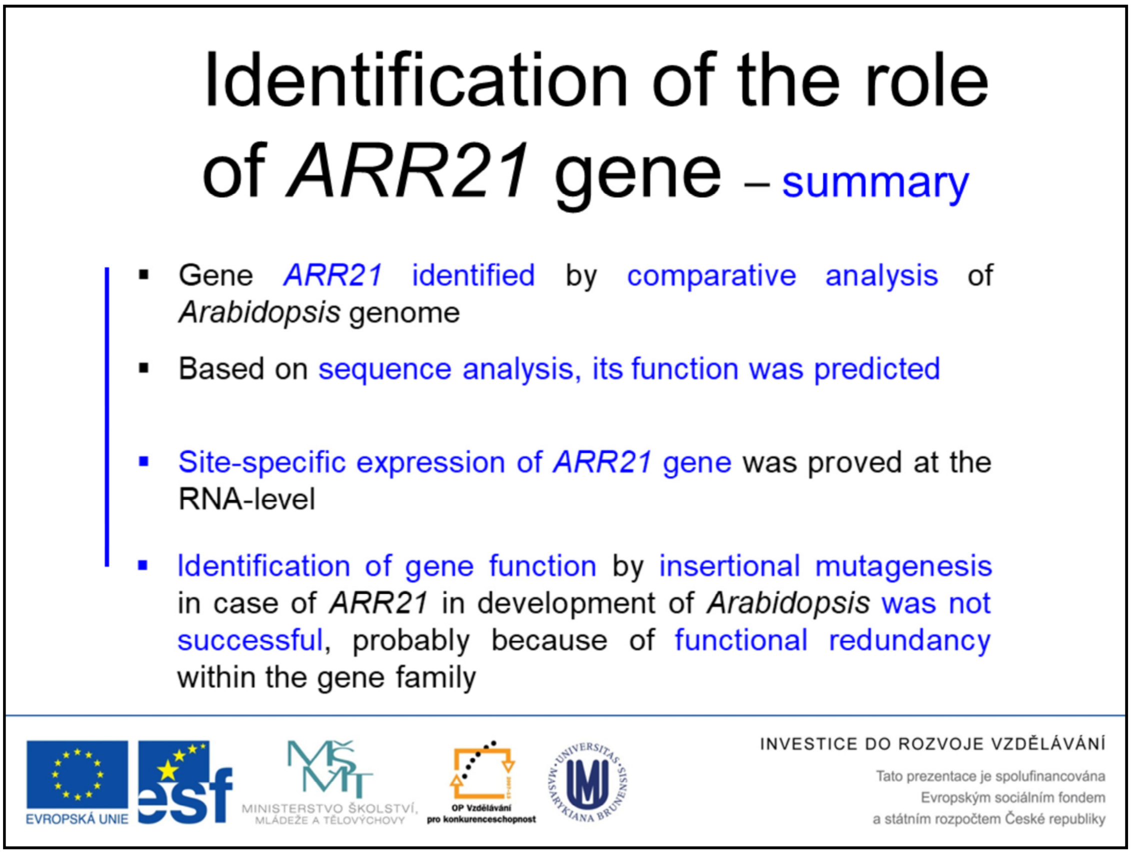


Identification of the role of *ARR21* gene – causes of absence of the phenotype

- Functional redundance within the gene family?
- Phenotype only under specific conditions

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

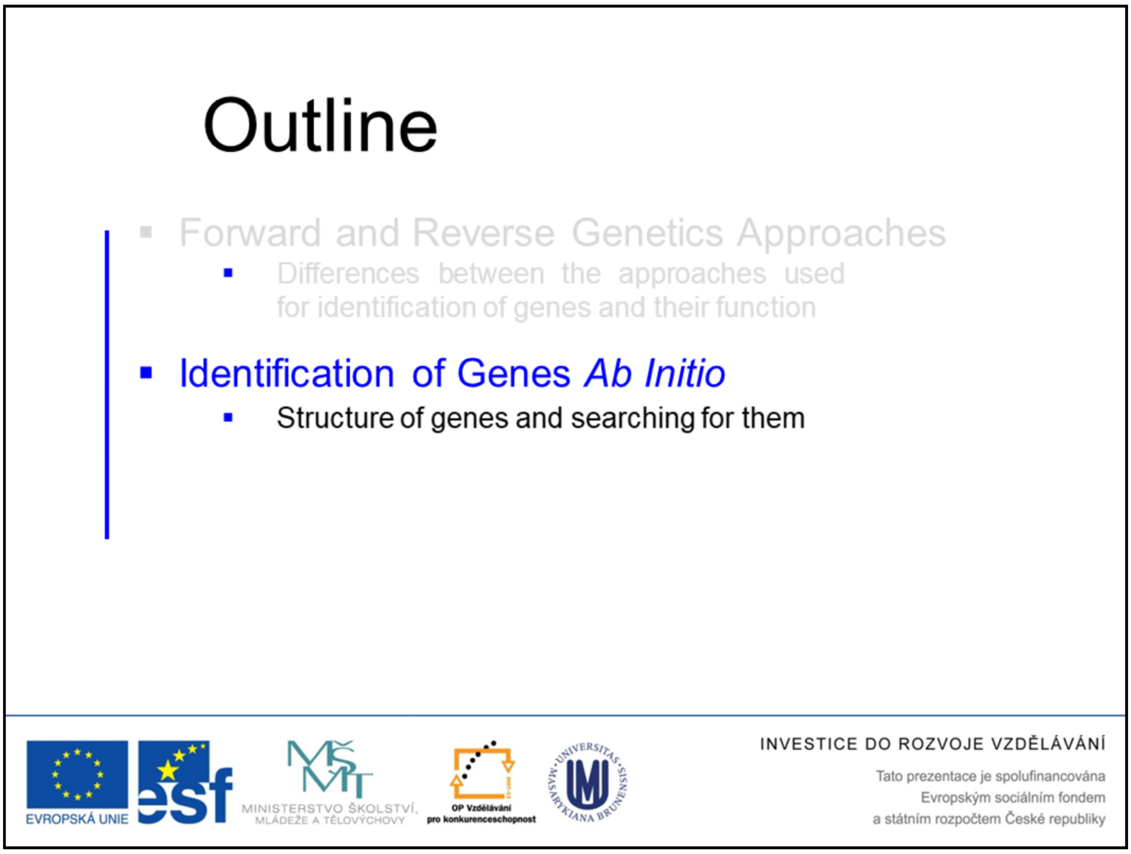
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Identification of the role of *ARR21* gene – summary

- Gene *ARR21* identified by comparative analysis of *Arabidopsis* genome
- Based on sequence analysis, its function was predicted
- Site-specific expression of *ARR21* gene was proved at the RNA-level
- Identification of gene function by insertional mutagenesis in case of *ARR21* in development of *Arabidopsis* was not successful, probably because of functional redundancy within the gene family

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

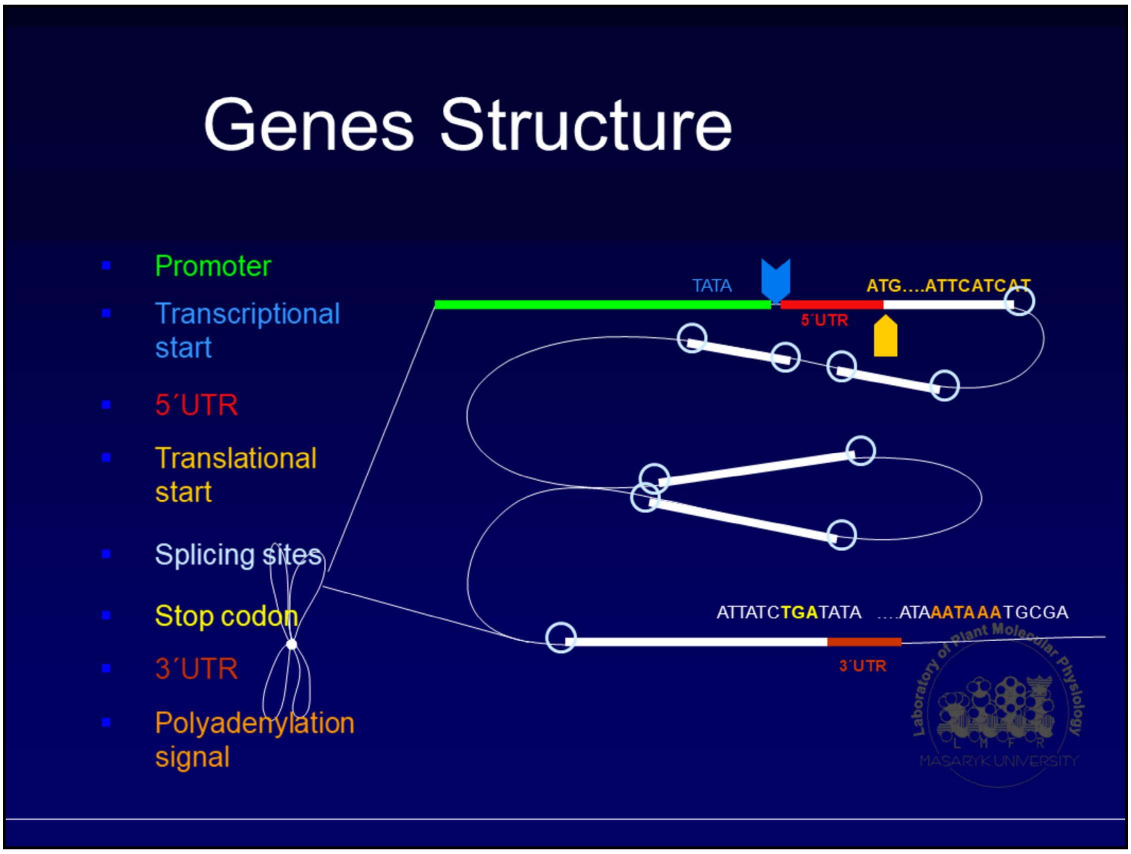


Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

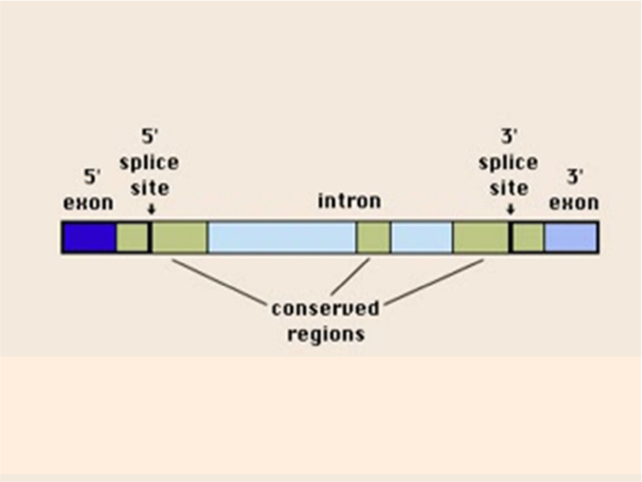
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



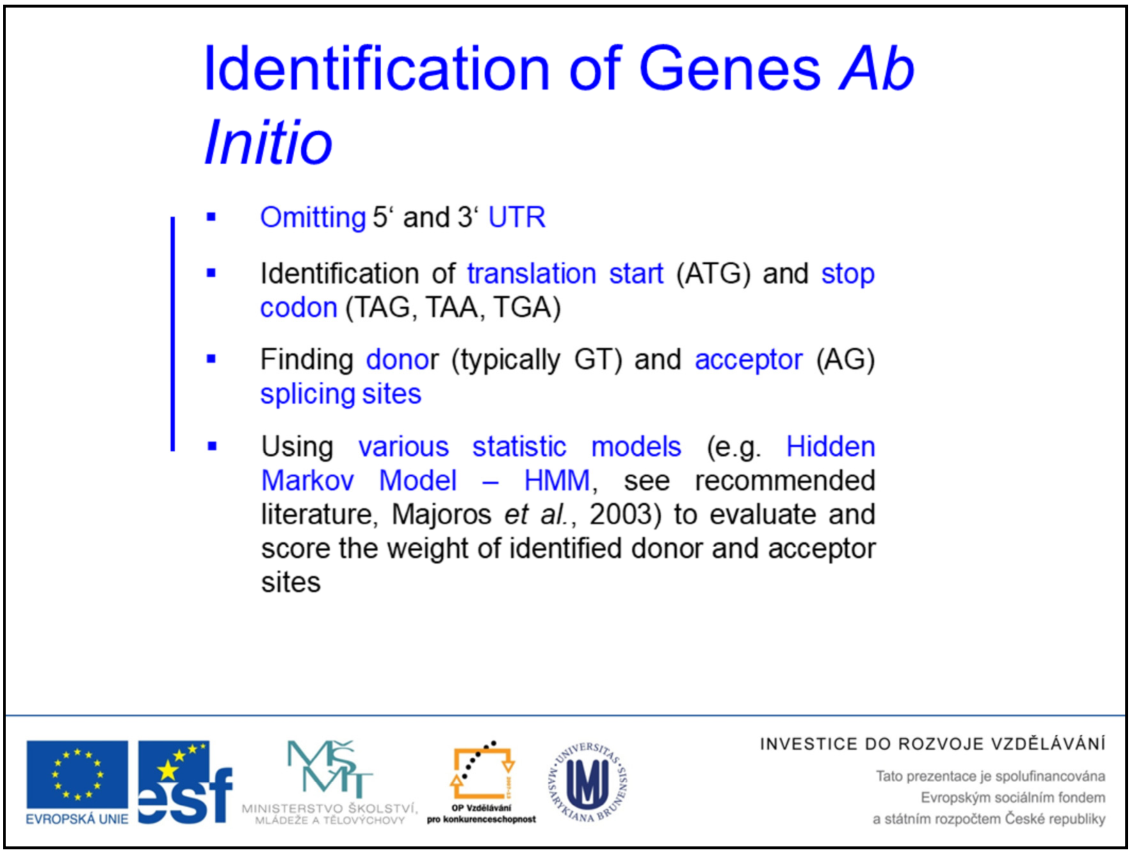
Genes Structure

- Promoter
- Transcriptional start
- 5´UTR
- Translational start
- Splicing sites
- Stop codon
- 3´UTR
- Polyadenylation signal

RNA Splicing



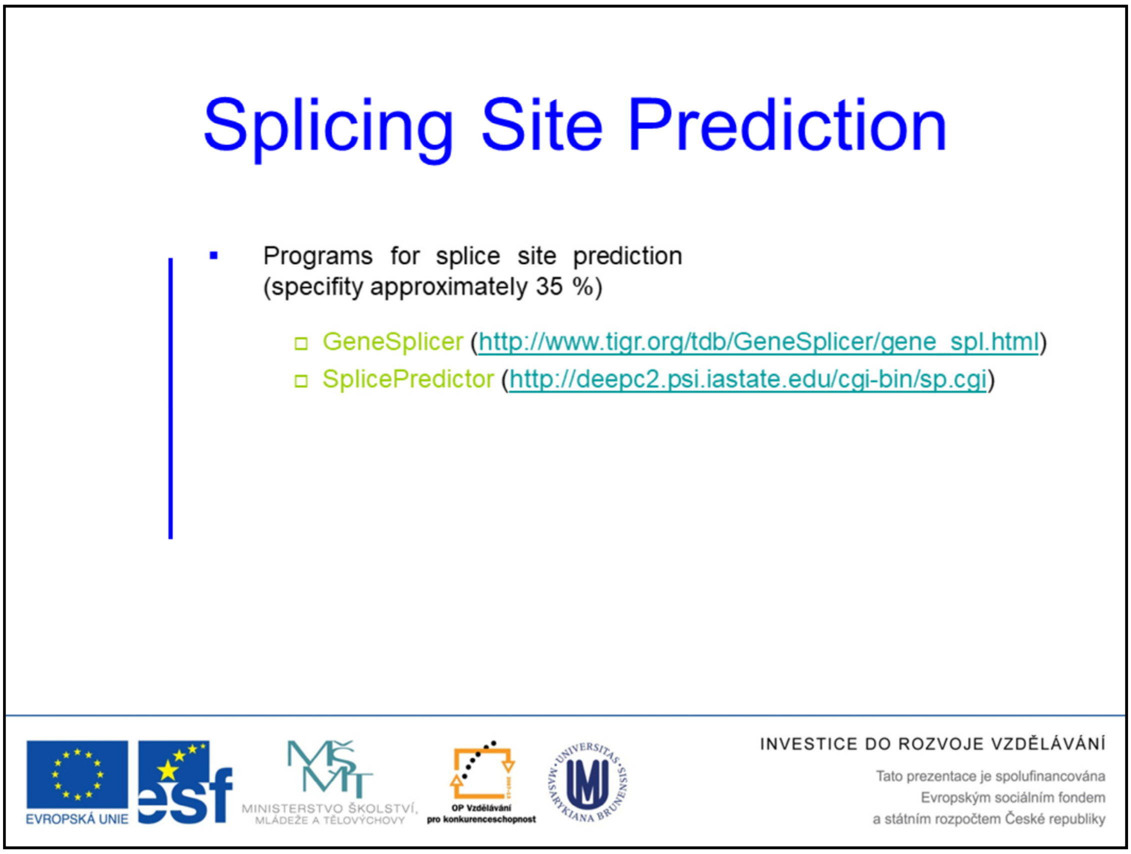
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Identification of Genes *Ab Initio*

- Omitting 5‘ and 3‘ UTR
- Identification of translation start (ATG) and stop codon (TAG, TAA, TGA)
- Finding donor (typically GT) and acceptor (AG) splicing sites
- Using various statistic models (e.g. Hidden Markov Model – HMM, see recommended literature, Majoros *et al.*, 2003) to evaluate and score the weight of identified donor and acceptor sites

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

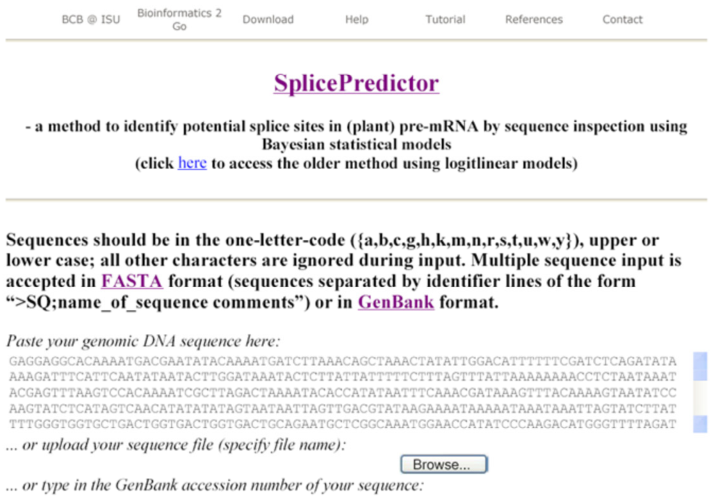


Splicing Site Prediction

- Programs for splice site prediction (specifity approximately 35 %)
  - GeneSplicer (http://www.tigr.org/tdb/GeneSplicer/gene_spl.html)
  - SplicePredictor (http://deepc2.psi.iastate.edu/cgi-bin/sp.cgi)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

SplicePredictor



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
AAAGATTTCATTCAATATAATACTTGGATAAATACTCTTATTATTTTTCTTTAGTTTATTAAAAAAAACCTCTAATAAAT
ACGAGTTTAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAAACGATAAAGTTTACAAAAGTAATATCC
AAGTATCTCATAGTCAACATATATATAGTAATAATTAGTTGACGTATAAGAAAATAAAAATAAATAAATTAGTATCTTAT
TTTGGGTGGTGCTGACTGGTGACTGGTGACTGCAGAATGCTCGGCAAATGGAACCATATCCCAAGACATGGGTTTTAGAT
```

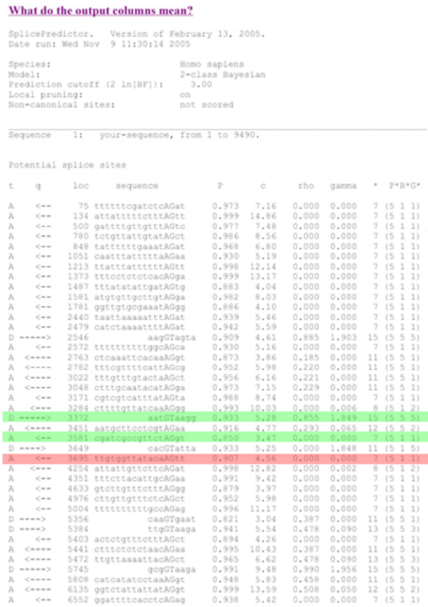
... or upload your sequence file (specify file name):

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

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

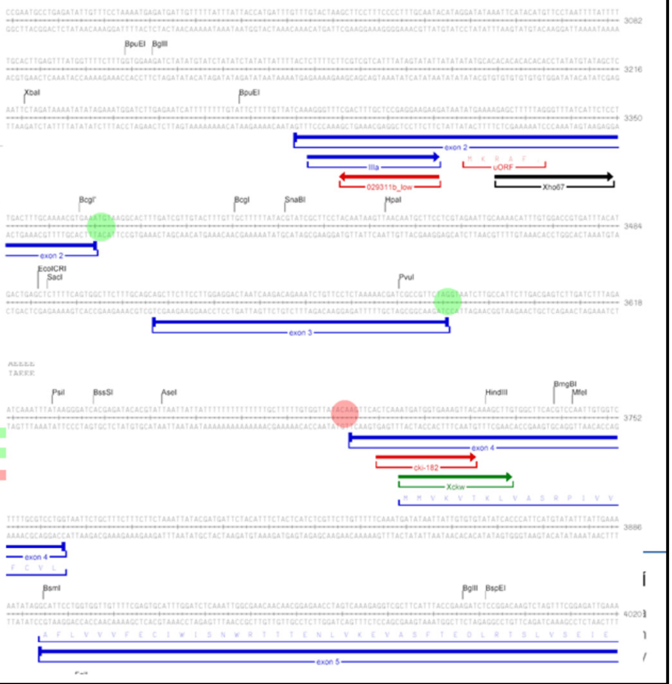
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

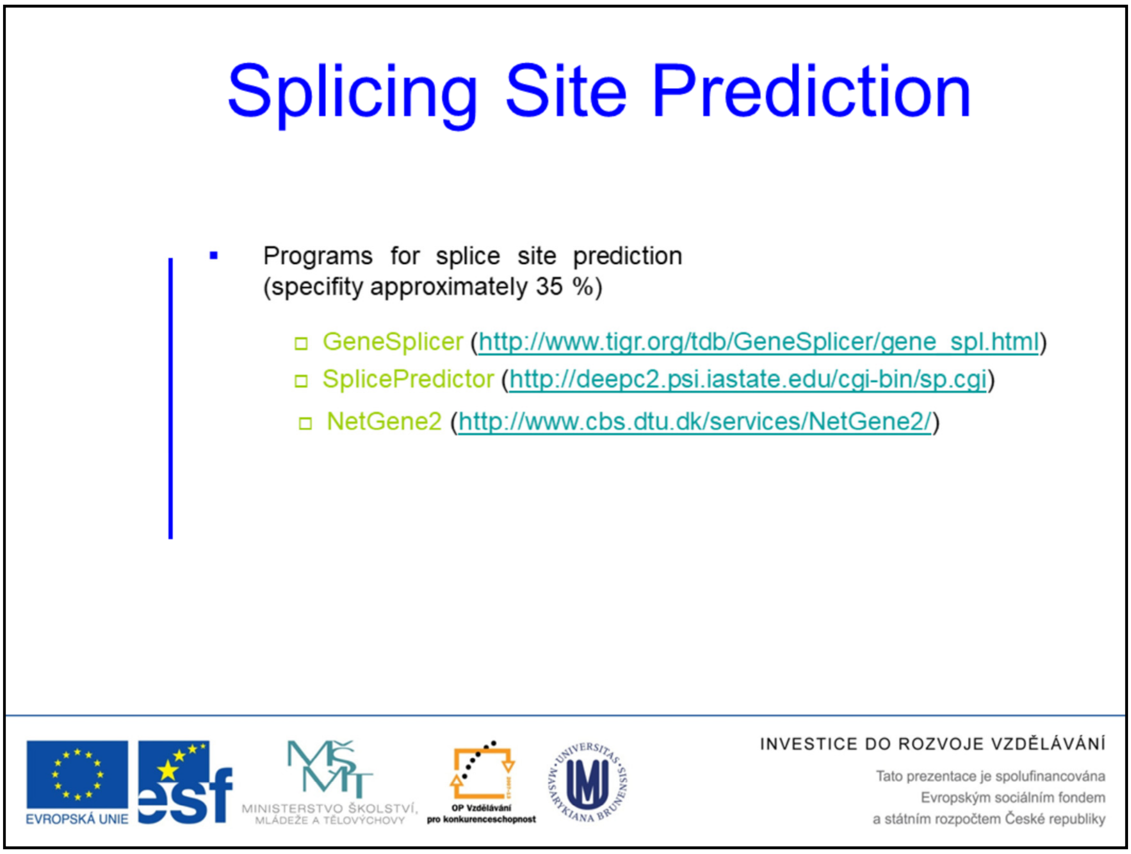
SplicePredictor



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 gattttgttgtttAGtc    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 ttatttatttttttAGtt   0.998  12.14   0.000  0.000    7 (5 1 1)
A    <--   1373 tttcctctctcacAGga    0.999  13.17   0.000  0.000    7 (5 1 1)
A    <--   1487 tttatatattgatAGtg    0.883   4.04   0.000  0.000    7 (5 1 1)
A    <--   1581 atgtgttgcttgtAGga    0.982   8.03   0.000  0.000    7 (5 1 1)
A    <--   1781 ggttgtgcgaaatAGgg    0.886   4.10   0.000  0.000    7 (5 1 1)
A    <--   2440 taattaaaaatttAGat    0.939   5.46   0.000  0.000    7 (5 1 1)
A    <--   2479 catctaaaattttAGat    0.942   5.59   0.000  0.000    7 (5 1 1)
D ----->   2546            aagGTagta  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 ctcaaattcacaaAGgt    0.873   3.86   0.185  0.000   11 (5 5 1)
A  <----   2782 tttcgttttcattAGcg    0.952   5.98   0.220  0.000   11 (5 5 1)
A  <----   3022 tttgtttgtactaAGct    0.956   6.16   0.221  0.000   11 (5 5 1)
A  <----   3048 ctttgcaatacatAGga    0.973   7.15   0.229  0.000   11 (5 5 1)
A    <--   3171 cgtcgtcatttatAGta    0.988   8.74   0.000  0.000    7 (5 1 1)
A   <---   3284 cttttgttatcaaAGgg    0.993  10.03   0.000  0.006    8 (5 1 2)
D ----->   3372            aatGTaagg  0.933   5.28   0.855  1.849   15 (5 5 5)
A  <----   3451 aatgcttcctcgtAGaa    0.916   4.77   0.293  0.065   12 (5 5 2)
A    <--   3581 cgatcgccgttctAGgt    0.850   3.47   0.000  0.000    7 (5 1 1)
D ---->    3649            cacGTatta  0.933   5.25   0.000  1.848   11 (5 1 5)
A    <--   3695 ttgtggttatacaAGtt    0.907   4.56   0.000  0.000    7 (5 1 1)
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 gtcttgtttctttAGgg    0.879   3.97   0.000  0.000    7 (5 1 1)
A    <--   4976 cttgttgtttctcAGct    0.952   5.98   0.000  0.000    7 (5 1 1)
A    <--   5004 tttttttttgccAGag     0.996  11.17   0.000  0.000    7 (5 1 1)
D ---->    5356            caaGTgaat  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 actctgtttctttAGct    0.894   4.26   0.000  0.000    7 (5 1 1)
A  <----   5441 ctttctctctaacAGaa    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 ggtctattattatAGgt    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)
```



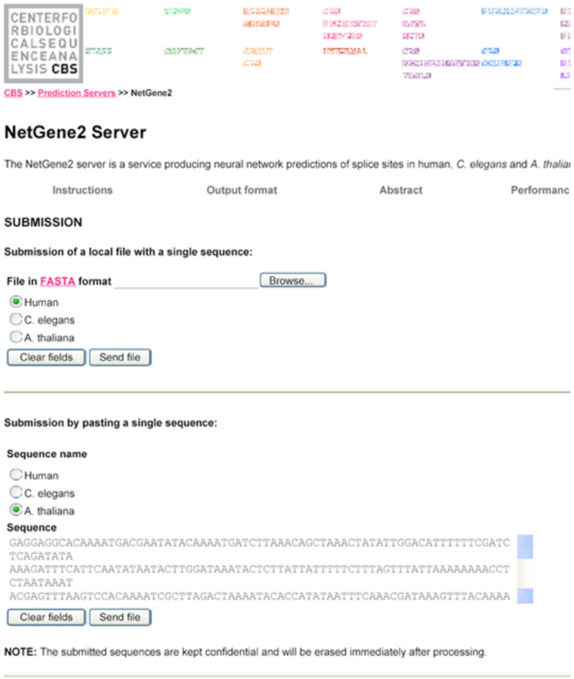


Splicing Site Prediction

- Programs for splice site prediction (specifity approximately 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/)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

NetGene2



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. thalia*

SUBMISSION

Submission of a local file with a single sequence:

File in FASTA format Browse...

● Human
○ C. elegans
○ A. thaliana

Submission by pasting a single sequence:

Sequence name

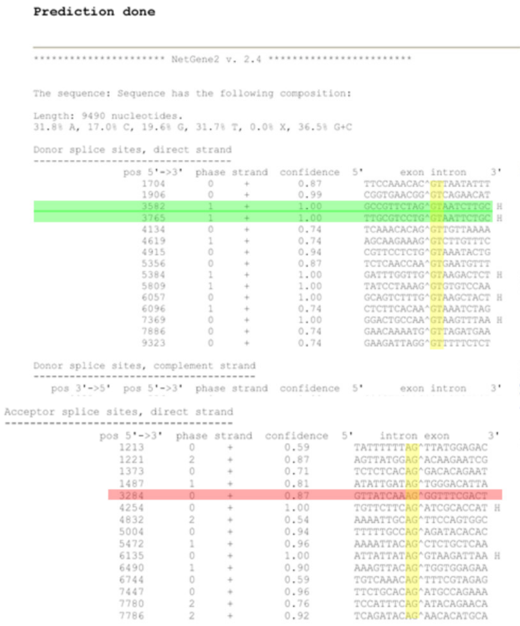
○ Human
○ C. elegans
● A. thaliana

Sequence

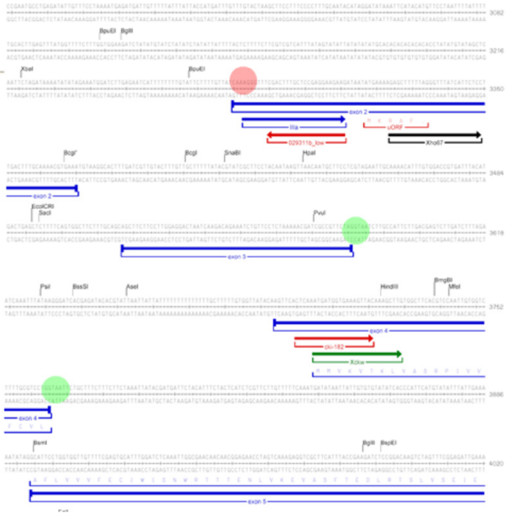
```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTTCGATC
TCAGATATA
AAAGATTTCATTCAATATAATACTTGGATAAATACTCTTATTATTTTTCTTTAGTTTATTAAAAAAAACCT
CTAATAAAT
ACGAGTTTAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAAACGATAAAGTTTACAAAA
```

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

NetGene2

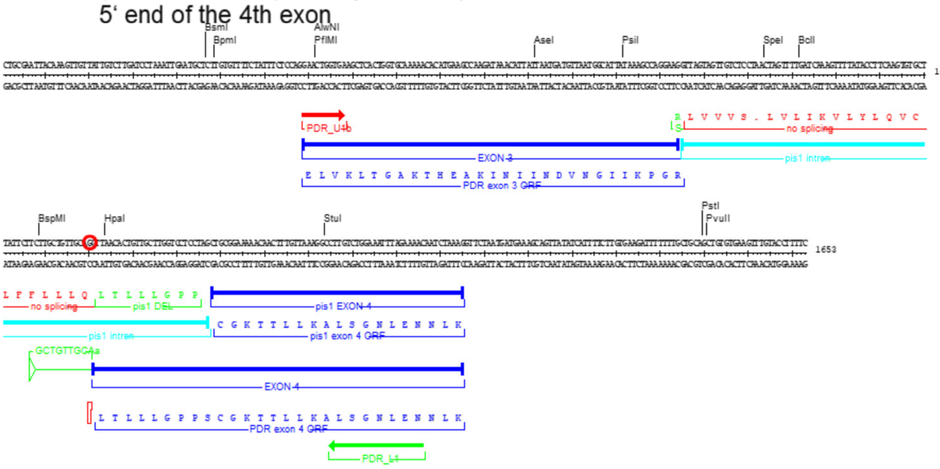


```
Prediction done

********************** 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     TTCCAAACAC^GTTAATATTT
                  1906       0      +       0.99     CGGTGAACGG^GTCAGAACAT
                  3582       1      +       1.00     GCCGTTCTAG^GTAATCTTGC H
                  3765       1      +       1.00     TTGCGTCCTG^GTAATTCTGC H
                  4134       0      +       0.74     TCAAACACAG^GTTGTTAAAA
                  4619       1      +       0.74     AGCAAGAAAG^GTCTTGTTTC
                  4915       0      +       0.94     CGTTCCTCTG^GTAAATACTG
                  5356       0      +       0.87     TCTCAACCAA^GTGAATGTTT
                  5384       1      +       1.00     GATTTGGTTG^GTAAGACTCT H
                  5809       1      +       1.00     TATCCTAAAG^GTGTGTCCAA
                  6057       0      +       1.00     GCAGTCTTTG^GTAAGCTACT H
                  6096       1      +       0.74     CTCTTCACAA^GTAAATCTAG
                  7369       0      +       1.00     GGACTGCCAA^GTAAGTTTAA H
                  7886       0      +       0.74     GAACAAAATG^GTTAGATGAA
                  9323       0      +       0.74     GAAGATTAGG^GTTTTTCTCT

Donor splice sites, complement strand
-------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'      exon intron    3'

Acceptor splice sites, direct strand
-------------------------------------
               pos 5'->3'  phase strand  confidence  5'    intron exon      3'
                  1213       0      +       0.59     TATTTTTTAG^TTATGGAGAC
                  1221       2      +       0.87     AGTTATGGAG^ACAAGAATCG
                  1373       0      +       0.71     TCTCTCACAG^GACACAGAAT
                  1487       1      +       0.81     ATATTGATAG^TGGGACATTA
                  3284       0      +       0.87     GTTATCAAAG^GGTTTCGACT
                  4254       0      +       1.00     TGTTCTTCAG^ATCGCACCAT H
                  4832       2      +       0.54     AAAATTGCAG^TTCCAGTGGC
                  5004       0      +       0.94     TTTTTGCCAG^AGATACACAC
                  5472       1      +       0.96     AAAATTACAG^CTCTGCTCAA
                  6135       0      +       1.00     ATTATTATAG^GTAAGATTAA H
                  6490       1      +       0.90     AAAGTTACAG^TGGTGGAGAA
                  6744       0      +       0.59     TGTCAAACAG^TTTCGTAGAG
                  7447       0      +       0.96     TTCTGCACAG^ATGCCAGAAA
                  7780       2      +       0.76     TCCATTTCAG^ATACAGAACA
                  7786       2      +       0.92     TCAGATACAG^AACACATGCA
```



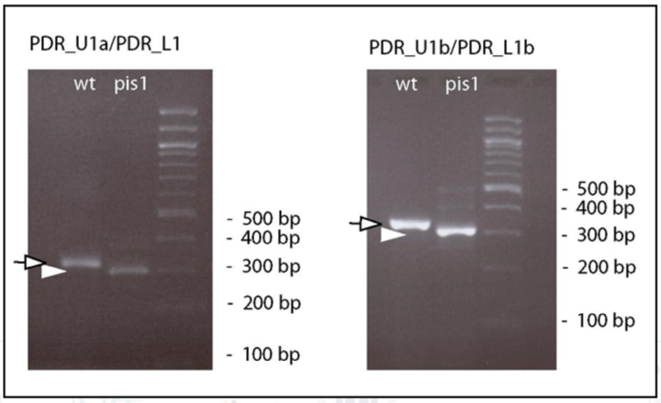
RNA Splicing and Adaptation

- Flexibility in splicing site recognition in plants in practice – example of developmental plasticity of (not only) plants
  - Identification of mutant with point mutation (transition G→A) exactly at the splice site at the 5‘ end of the 4th exon



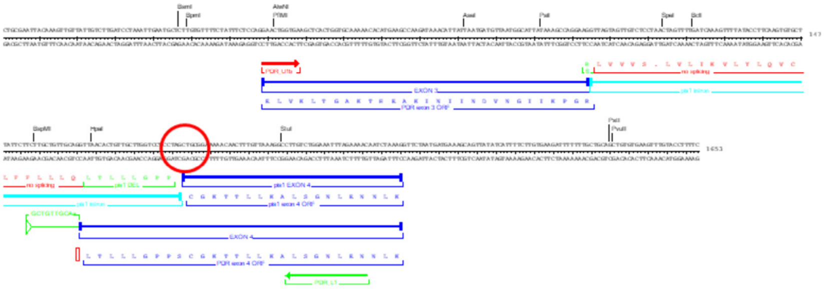
RNA Splicing and Adaptation

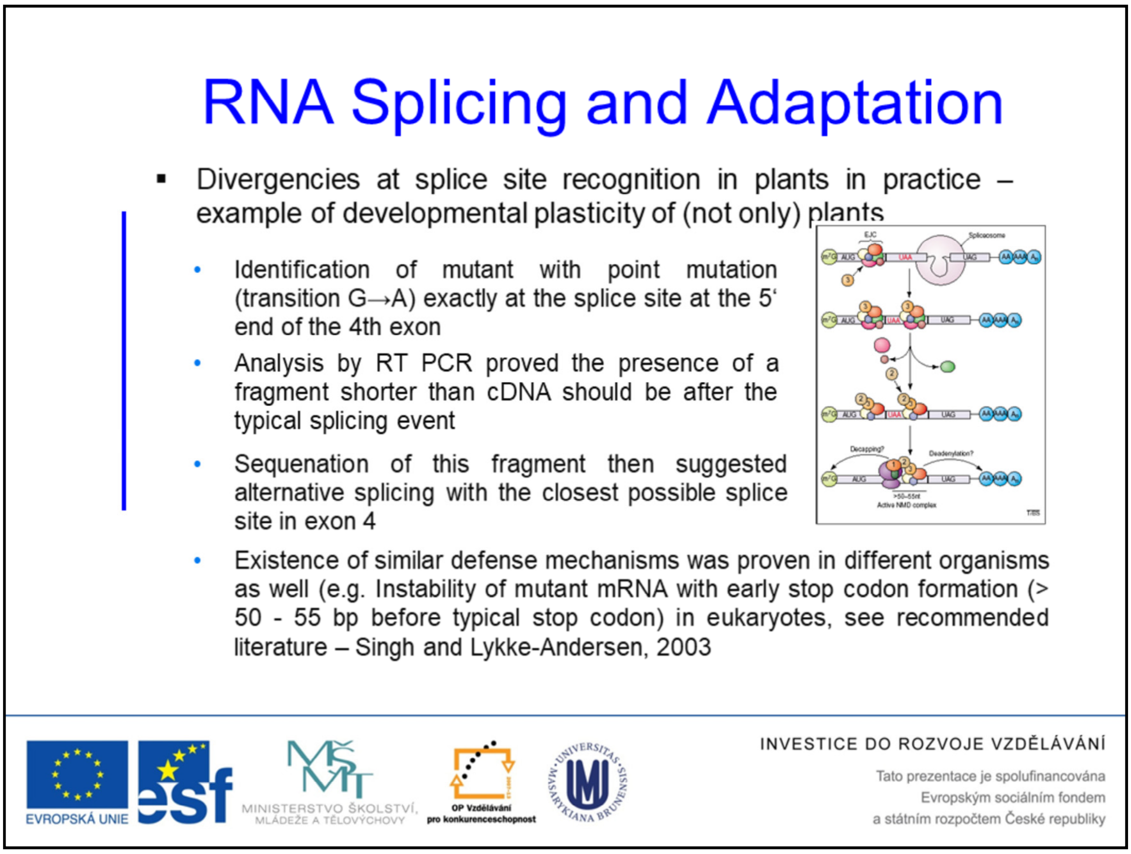
- Identification of mutant with point mutation (transition G→A) exactly at the splice site at the 5‘ end of the 4th exon
- Analysis by RT PCR proved the presence of a fragment shorter than cDNA should be after the typical splicing event



RNA Splicing and Adaptation

- Flexibility in splicing site recognition in plants in practice – example of developmental plasticity of (not only) plants
  - Identification of mutant with point mutation (transition G→A) exactly at the splice site at the 5‘ end of the 4th exon
  - Analysis by RT PCR proved the presence of a fragment shorter than cDNA should be after the typical splicing event
  - Sequenation of this fragment then suggested alternative splicing with the closest possible splice site in exon 4





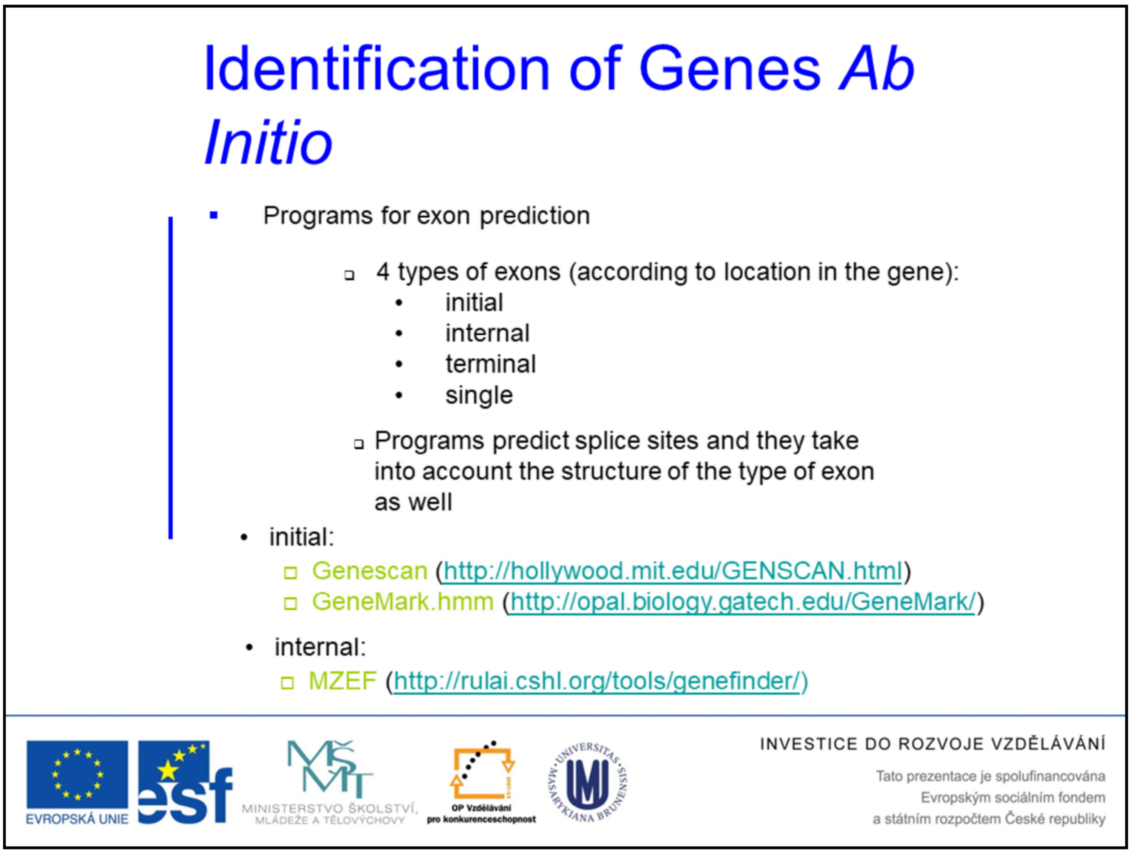
RNA Splicing and Adaptation

- Divergencies at splice site recognition in plants in practice – example of developmental plasticity of (not only) plants
  - Identification of mutant with point mutation (transition G→A) exactly at the splice site at the 5‘ end of the 4th exon
  - Analysis by RT PCR proved the presence of a fragment shorter than cDNA should be after the typical splicing event
  - Sequenation of this fragment then suggested alternative splicing with the closest possible splice site in exon 4
  - Existence of similar defense mechanisms was proven in different organisms as well (e.g. Instability of mutant mRNA with early stop codon formation (> 50 - 55 bp before typical stop codon) in eukaryotes, see recommended literature – Singh and Lykke-Andersen, 2003

EJC
Spliceosome
m7G
AUG
UAA
UAG
AA
AAA
An
3
2
1
Decapping?
Deadenylation?
>50–55nt
Active NMD complex
TiBS
EVROPSKÁ UNIE
esf
MŠMT
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



Identification of Genes *Ab Initio*

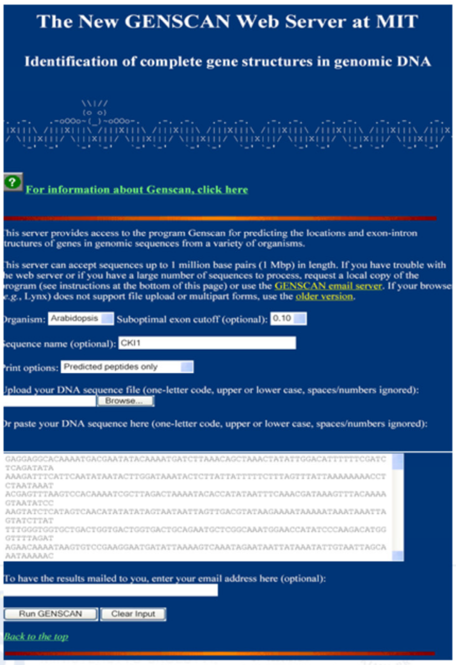
- Programs for exon prediction
  - ❑ 4 types of exons (according to location in the gene):
    - initial
    - internal
    - terminal
    - single
  - ❑ Programs predict splice sites and they take into account the structure of the type of exon as well
- initial:
  - □ Genescan (http://hollywood.mit.edu/GENSCAN.html)
  - □ GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/)
- internal:
  - □ MZEF (http://rulai.cshl.org/tools/genefinder/)

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

❑programy kromě rozpoznávání míst sestřihu zohledňují i strukturu jednotlivých typů exonů

GENSCAN



The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA

For information about Genscan, click here

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

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

)rganism: Arabidopsis Suboptimal exon cutoff (optional): 0.10

equence name (optional): CKI1

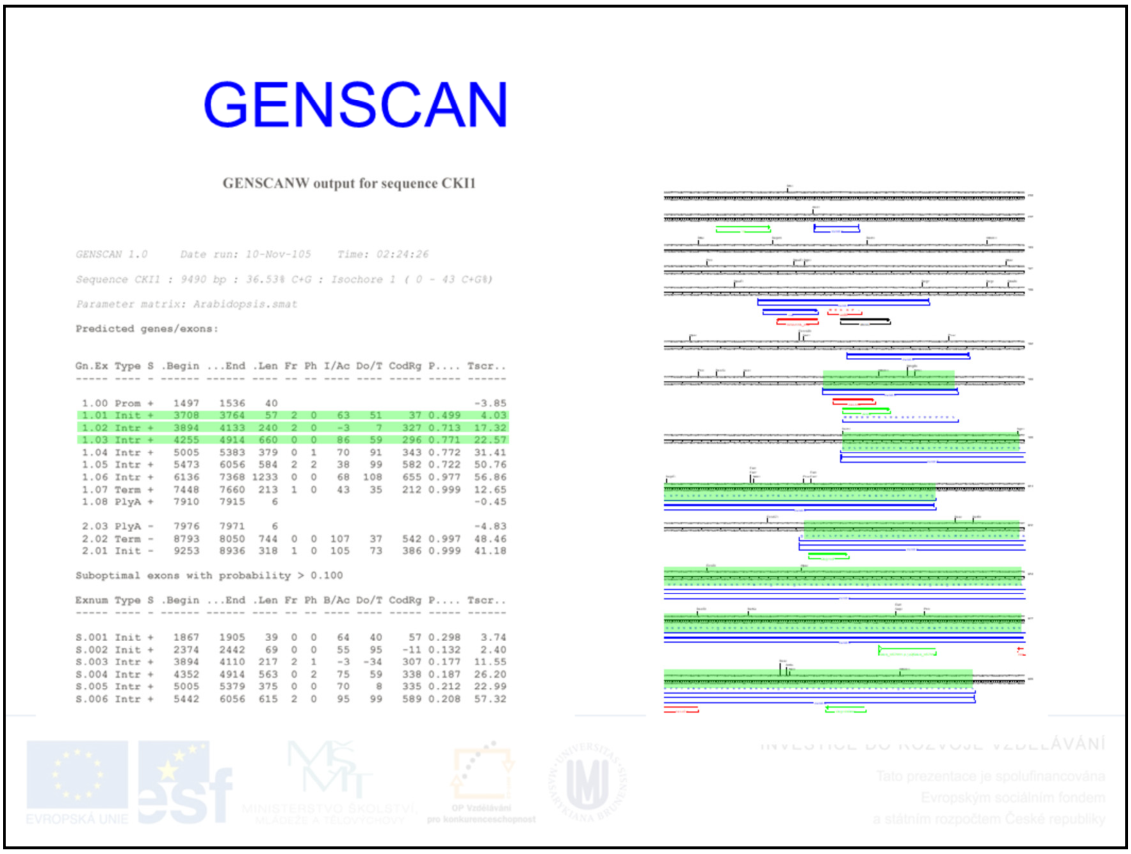
rint options: Predicted peptides only

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

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

```
GAGGAGGCACAAAATGACGAATATACAAAATGATCTTAAACAGCTAAACTATATTGGACATTTTTTCGATC
TCAGATATA
AAAGATTTCATTCAATATAATACTTGGATAAATACTCTTATTATTTTTCTTTAGTTTATTAAAAAAAACCT
CTAATAAAT
ACGAGTTTAAGTCCACAAAATCGCTTAGACTAAAATACACCATATAATTTCAAACGATAAAGTTTACAAAA
GTAATATCC
AAGTATCTCATAGTCAACATATATATAGTAATAATTAGTTGACGTATAAGAAAATAAAAATAAATAAATTA
GTATCTTAT
TTTGGGTGGTGCTGACTGGTGACTGGTGACTGCAGAATGCTCGGCAAATGGAACCATATCCCAAGACATGG
GTTTTAGAT
AGAACAAAATAAGTGTCCGAAGGAATGATATTAAAAGTCAAATAGAATAATTATAAATATTGTAATTAGCA
AATAAAAAC
```

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



GENSCAN

GENSCANW output for sequence CKI1

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

Sequence CKI1 : 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
 1.01 Init +   3708   3764   57  2  0   63   51    37 0.499   4.03
 1.02 Intr +   3894   4133  240  2  0   -3    7   327 0.713  17.32
 1.03 Intr +   4255   4914  660  0  0   86   59   296 0.771  22.57
 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
```

Explanation Gn.Ex : gene number, exon number (for reference) **Type** : Init = Initial exon (ATG to 5' splice site) Intr = Internal exon (3' splice site to 5' splice site) Term = Terminal exon (3' splice site to stop codon) Sngl = Single-exon gene (ATG to stop) Prom = Promoter (TATA box / initation site) PlyA = poly-A signal (consensus: AATAAA) **S** : DNA strand (+ = input strand; - = opposite strand) **Begin** : beginning of exon or signal (numbered on input strand) **End** : end point of exon or signal (numbered on input strand) **Len** : length of exon or signal (bp) **Fr** : reading frame (a forward strand codon ending at x has frame x mod 3). For example, if nucleotides 1,2,3 of the sequence are read as a codon, that's called reading frame 0. If 2,3,4 are read as a codon, that's reading frame 1. If 3,4,5 are read as a codon, that's reading frame 2, and so on. This information, together with the starting and ending positions of the exon, is sufficient to give the amino acid sequence encoded by the exon. Another use of the reading frame is that if you see two adjacent predicted exons separated by a relatively short intron which share the same reading frame, it may be worth looking at the possibility that the intervening intron is not correct, i.e. that the two exons plus the intervening intron might form one long exon (assuming there are no inframe stops in the intron, of course). **Ph** : net phase of exon (exon length modulo 3). For example, an exon of length 15 bp has net phase 0 since 15 is divisible by 3, an exon of length 16 bp has net phase 1 because 16 divided by 3 leaves a remainder of 1, an exon of length 17 bp has net phase 2, and an exon of length 18 bp has net phase 0 again. The point of this is that exons whose net phase is 0 can be omitted from the gene without disrupting the reading frame: such exons are candidates for being either 1) incorrect, or 2) alternatively spliced. **I/Ac** : initiation signal or 3' splice site score (tenth bit units; x 10). If below zero, probably not a real acceptor site. **Do/T** : 5' splice site or termination signal score (tenth bit units; x 10) If below zero, probably not a real donor site. **CodRg** : coding region score (tenth bit units) **P** : probability of exon (sum over all parses containing exon). This quantity is close to the actual probability that the predicted exon is correct. **Tscr** : exon score (depends on length, I/Ac, Do/T and CodRg scores).

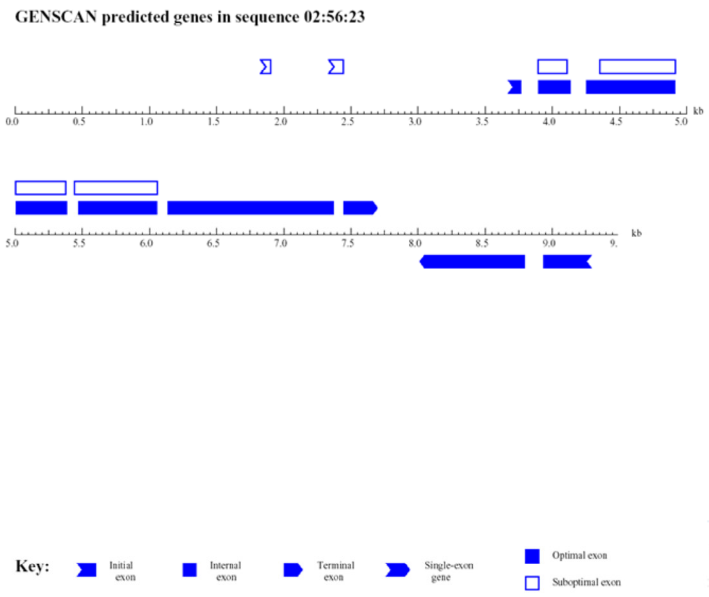
Comments The SCORE of a predicted feature (e.g., exon or splice site) is a log-odds measure of the quality of the feature based on local sequence properties. For example, a predicted 5' splice site with score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor (more than likely not a real donor site). The PROBABILITY of a predicted exon is the estimated probability under GENSCAN's model of genomic sequence structure that the exon is correct. This probability depends in general on global as well as local sequence properties, e.g., it depends on how well the exon fits with neighboring exons. It has been shown that predicted exons with higher probabilities are more likely to be correct than those with lower probabilities.

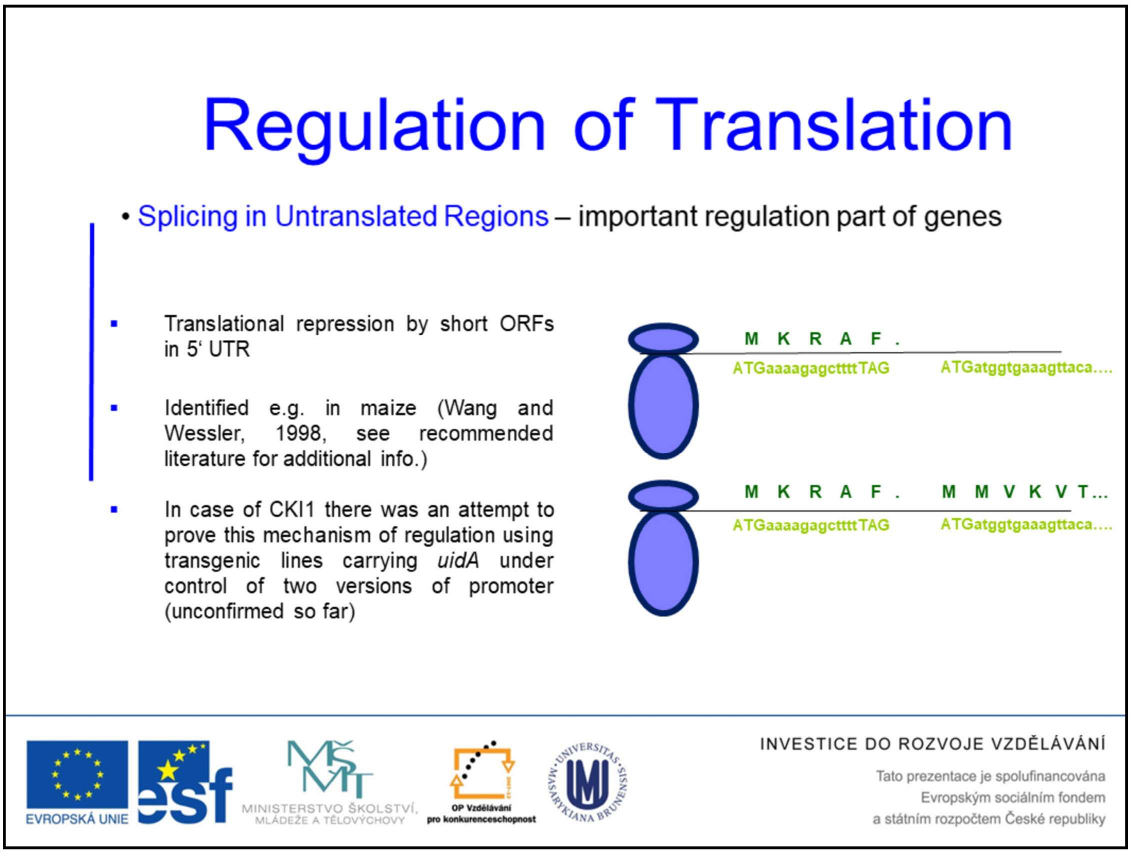
What are the suboptimal exons?

Under the probabilistic model of gene structural and compositional properties used by GENSCAN, each possible "parse" (gene structure description) which is compatible with the sequence is assigned a probability. The default output of the program is simply the "optimal" (highest probability) parse of the sequence. The exons in this optimal parse are referred to as "optimal exons" and the translation products of the corresponding "optimal genes" are printed as GENSCAN predicted peptides. (All the data in our J Mol Biol paper and on the other GENSCAN web pages refer exclusively to the optimal parse/optimal exons.) Of course, the optimal parse does not always correspond to the actual (biological) parse of the sequence, that is, the actual set of exons/genes present. In addition, there may be more than one parse which can be considered "correct", for example, in the case of a gene which is alternatively transcribed, translated or spliced. For both of these reasons, it may be of interest to consider "suboptimal" ("near-optimal") exons as well, i.e. exons which have reasonably high probability but are not present in the optimal parse. Specifically, for every potential exon E in the sequence, the probability P(E) is defined as the sum of the probabilities under the model of all possible "parses" (gene structures) which contain the exact exon E in the correct reading frame. (This quantity is calculated as described on the GENSCAN exon probability page.) Given a probability cutoff C, suboptimal exons are those potential exons with P(E) > C which are not present in the optimal parse.

Suboptimal exons have a variety of potential uses. First, suboptimal exons sometimes correspond to real exons which were missed for whatever reason by the optimal parse of the sequence. Second, regions of a prediction which contain multiple overlapping and/or incompatible optimal and suboptimal exons may in some cases indicate alternatively spliced regions of a gene (Burge & Karlin, in preparation). The probability cutoff C used to determine which potential exons qualify as suboptimal exons can be set to any of a range of values between 0.01 and 1.00. The default value on the web page is 1.00, meaning that no suboptimal exons are printed. For most applications, a cutoff value of about 0.10 is recommended. Setting the value much lower than 0.10 will often lead to an explosion in the number of suboptimal exons, most of which will probably not be useful. On the other hand, if the value is set much higher than 0.10, then potentially interesting suboptimal exons may be missed.

GENSCAN


GENSCAN predicted genes in sequence 02:56:23
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
kb
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.
kb
Key:
Initial exon
Internal exon
Terminal exon
Single-exon gene
Optimal exon
Suboptimal exon



Regulation of Translation

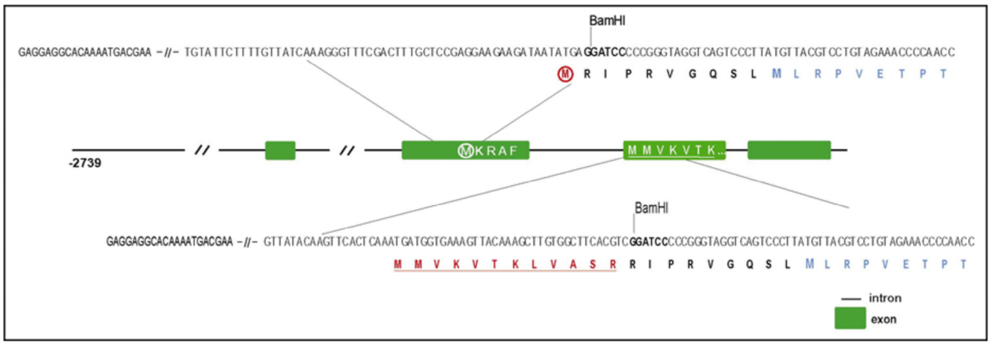
- Splicing in Untranslated Regions – important regulation part of genes

- Translational repression by short ORFs in 5‘ UTR
- Identified e.g. in maize (Wang and Wessler, 1998, see recommended literature for additional info.)
- In case of CKI1 there was an attempt to prove this mechanism of regulation using transgenic lines carrying *uidA* under control of two versions of promoter (unconfirmed so far)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Regulation of translation

- Functional purpose of splicing in untranslated regions – important regulation part of genes
  - In case of CKI1 there was an attempt to prove this mechanism of regulation using transgenic lines carrying *uidA* under control of two versions of promoter (unconfirmed so far)

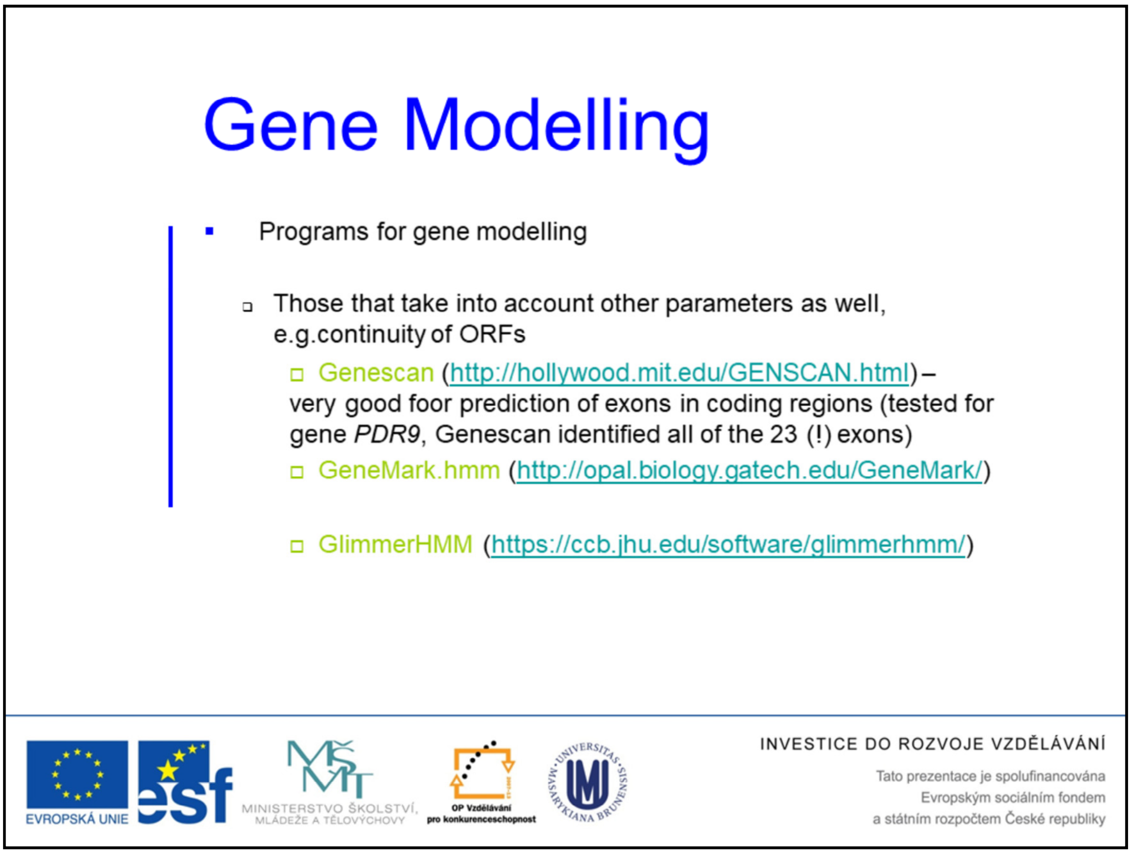


BamHI
GAGGAGGCACAAAATGACGAA –//– TGTATTCTTTTGTTATCAAAGGGTTTCGACTTTGCTCCGAGGAAGAAGATAATATGAGGATCCCCCGGGTAGGTCAGTCCCTTATGTTACGTCCTGTAGAAACCCCAACC
M R I P R V G Q S L M L R P V E T P T
-2739
// //
MKRAF
M M V K V T K…
BamHI
GAGGAGGCACAAAATGACGAA –//– GTTATACAAGTTCACTCAAATGATGGTGAAAGTTACAAAGCTTGTGGCTTCACGTCGGATCCCCCGGGTAGGTCAGTCCCTTATGTTACGTCCTGTAGAAACCCCAACC
M M V K V T K L V A S R R I P R V G Q S L M L R P V E T P T
intron
exon

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost



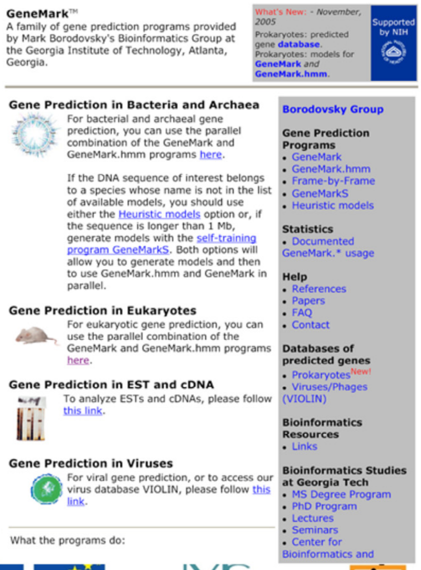
Gene Modelling

- Programs for gene modelling
  - Those that take into account other parameters as well, e.g.continuity of ORFs
    - Genescan (http://hollywood.mit.edu/GENSCAN.html) – very good foor prediction of exons in coding regions (tested for gene *PDR9*, Genescan identified all of the 23 (!) exons)
    - GeneMark.hmm (http://opal.biology.gatech.edu/GeneMark/)
    - GlimmerHMM (https://ccb.jhu.edu/software/glimmerhmm/)

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

GeneMark



GeneMark™
A family of gene prediction programs provided by Mark Borodovsky's Bioinformatics Group at the Georgia Institute of Technology, Atlanta, Georgia.

What's New: - *November, 2005*
Prokaryotes: predicted gene **database**.
Prokaryotes: models for **GeneMark** *and* **GeneMark.hmm**.

Supported by NIH

Gene Prediction in Bacteria and Archaea

For bacterial and archaeal gene prediction, you can use the parallel combination of the GeneMark and GeneMark.hmm programs here.

If the DNA sequence of interest belongs to a species whose name is not in the list of available models, you should use either the Heuristic models option or, if the sequence is longer than 1 Mb, generate models with the self-training program GeneMarkS. Both options will allow you to generate models and then to use GeneMark.hmm and GeneMark in parallel.

Gene Prediction in Eukaryotes

For eukaryotic gene prediction, you can use the parallel combination of the GeneMark and GeneMark.hmm programs here.

Gene Prediction in EST and cDNA

To analyze ESTs and cDNAs, please follow this link.

Gene Prediction in Viruses

For viral gene prediction, or to access our virus database VIOLIN, please follow this link.

What the programs do:

Borodovsky Group

Gene Prediction Programs
- GeneMark
- GeneMark.hmm
- Frame-by-Frame
- GeneMarkS
- Heuristic models

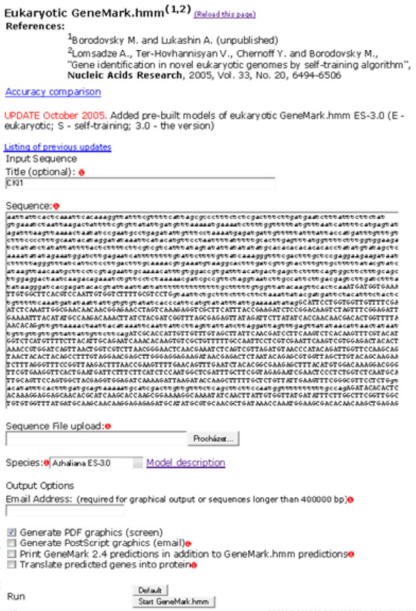
Statistics
- Documented GeneMark.* usage

Help
- References
- Papers
- FAQ
- Contact

Databases of predicted genes
- Prokaryotes New!
- Viruses/Phages (VIOLIN)

Bioinformatics Resources
- Links

Bioinformatics Studies at Georgia Tech
- MS Degree Program
- PhD Program
- Lectures
- Seminars
- Center for Bioinformatics and



Eukaryotic GeneMark.hmm(1,2) (Reload this page)

References:

[1] Borodovsky M. and Lukashin A. (unpublished)
[2] Lomsadze A., Ter-Hovhannisyan V., Chernoff Y. and Borodovsky M., "Gene identification in novel eukaryotic genomes by self-training algorithm", **Nucleic Acids Research**, 2005, Vol. 33, No. 20, 6494-6506

Accuracy comparison

UPDATE October 2005. Added pre-built models of eukaryotic GeneMark.hmm ES-3.0 (E - eukaryotic; S - self-training; 3.0 - the version)

Listing of previous updates

Input Sequence

Title (optional): CKI1

Sequence: [illegible]

Sequence File upload: Procházet...

Species: Athaliana ES-3.0 Model description

Output Options

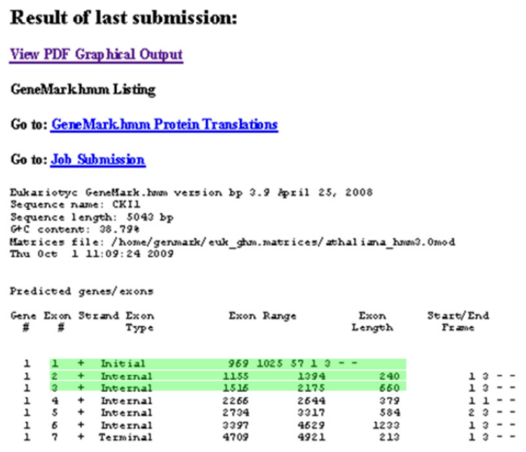
Email Address: (required for graphical output or sequences longer than 400000 bp)

☑ Generate PDF graphics (screen)
☐ Generate PostScript graphics (email)
☐ Print GeneMark 2.4 predictions in addition to GeneMark.hmm predictions
☐ Translate predicted genes into protein

Run

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
LÁVÁNÍ
...nancována Evropským sociálním fondem a státním rozpočtem České republiky

GeneMark



Result of last submission:

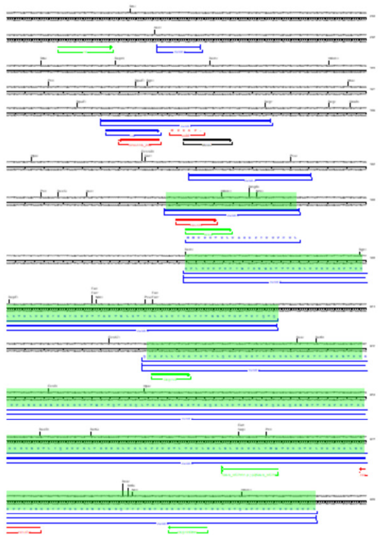
View PDF Graphical Output

GeneMark.hmm Listing

Go to: GeneMark.hmm Protein Translations

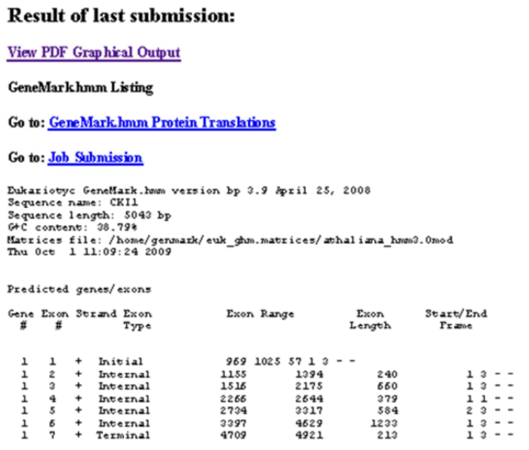
Go to: Job Submission

```
Eukariotyc GeneMark.hmm version bp 3.9 April 25, 2008
Sequence name: CKI1
Sequence length: 5043 bp
G+C content: 38.79%
Matrices file: /home/genmark/euk_ghm.matrices/athaliana_hmm3.0mod
Thu Oct  1 11:09:24 2009


Predicted genes/exons

Gene Exon Strand Exon            Exon Range          Exon     Start/End
  #    #          Type                              Length      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 - -
```



EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
/ZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

GeneMark



Result of last submission:

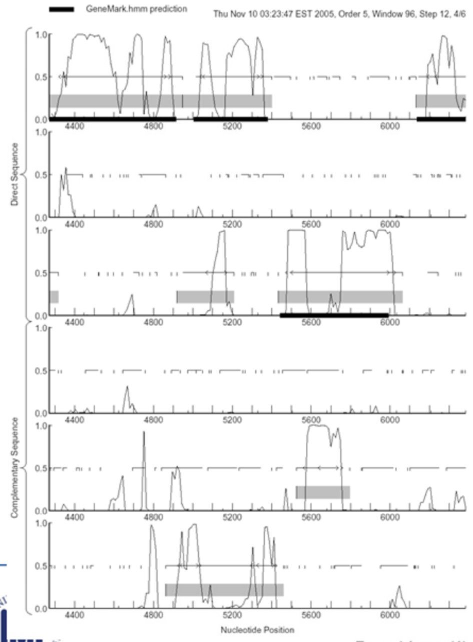
[View PDF Graphical Output]

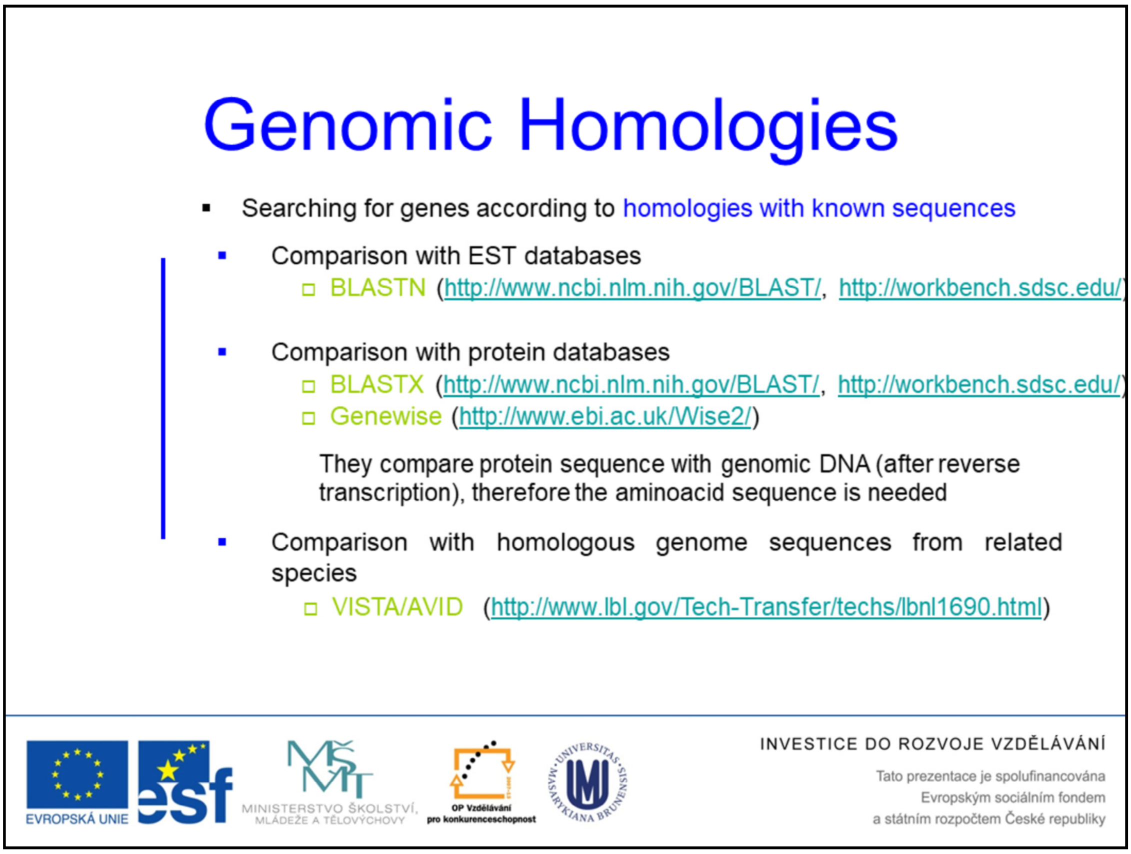
GeneMark.hmm Listing

Go to: [GeneMark.hmm Protein Translations]

Go to: [Job Submission]

```
Dukariotyc GeneMark.hmm version bp 3.9 April 25, 2008
Sequence name: CKI1
Sequence length: 5043 bp
G+C content: 38.79%
Matrices file: /home/genmark/euk_ghm.matrices/athaliana_hmm3.0mod
Thu Oct  1 11:09:24 2009


Predicted genes/exons

Gene Exon Strand Exon          Exon Range          Exon      Start/End
 #    #         Type                              Length       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 - -
```



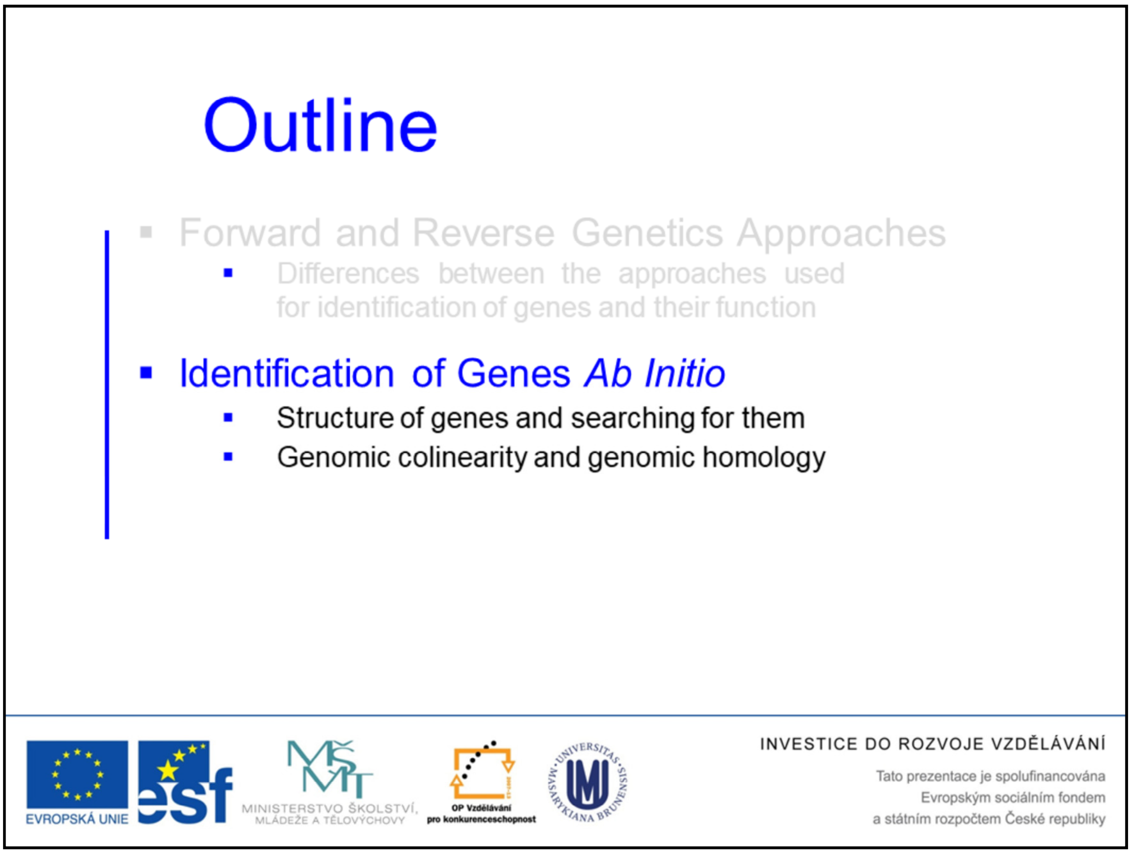


Genomic Homologies

- Searching for genes according to homologies with known sequences
  - Comparison with EST databases
    - □ BLASTN (http://www.ncbi.nlm.nih.gov/BLAST/, http://workbench.sdsc.edu/)
  - Comparison with protein databases
    - □ BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/, http://workbench.sdsc.edu/)
    - □ Genewise (http://www.ebi.ac.uk/Wise2/)

      They compare protein sequence with genomic DNA (after reverse transcription), therefore the aminoacid sequence is needed
  - Comparison with homologous genome sequences from related species
    - □ VISTA/AVID (http://www.lbl.gov/Tech-Transfer/techs/lbnl1690.html)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them
  - Genomic colinearity and genomic homology

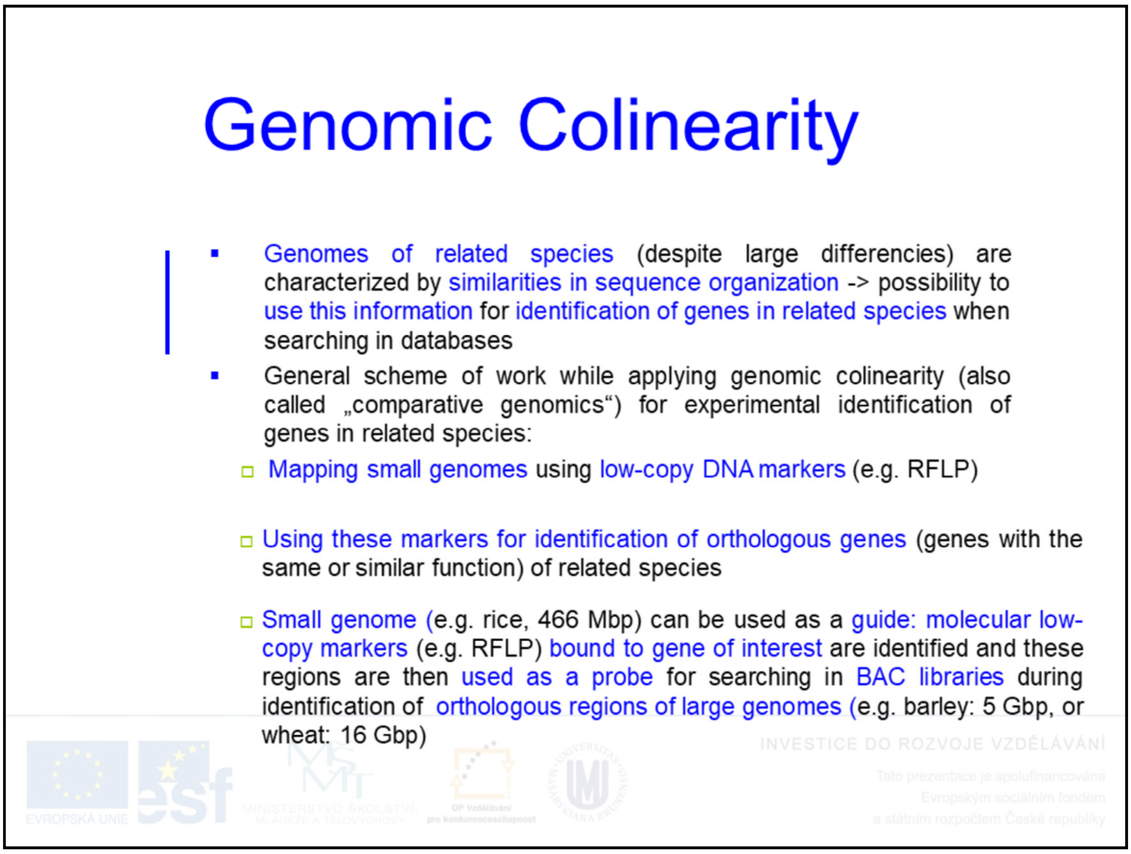
EVROPSKÁ UNIE

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

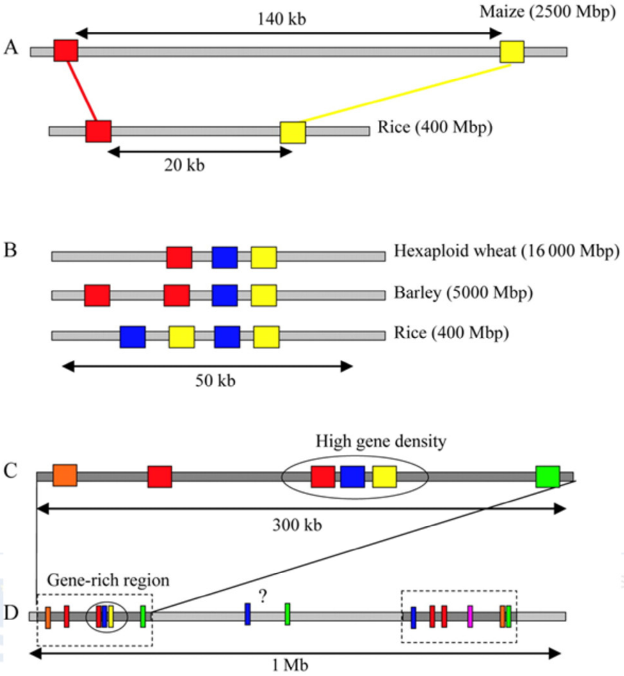


Genomic Colinearity

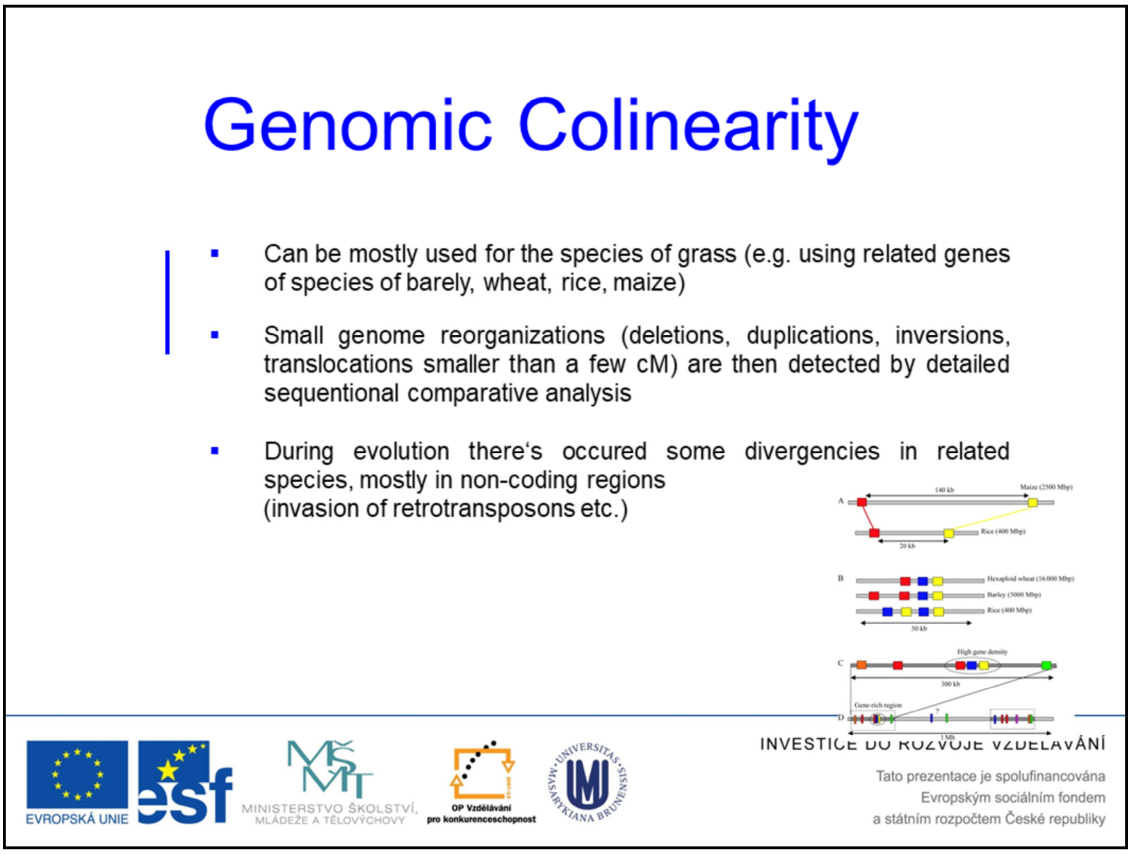
- Genomes of related species (despite large differencies) are characterized by similarities in sequence organization -> possibility to use this information for identification of genes in related species when searching in databases
- General scheme of work while applying genomic colinearity (also called „comparative genomics“) for experimental identification of genes in related species:
  - □ Mapping small genomes using low-copy DNA markers (e.g. RFLP)
  - □ Using these markers for identification of orthologous genes (genes with the same or similar function) of related species
  - □ Small genome (e.g. rice, 466 Mbp) can be used as a guide: molecular low-copy markers (e.g. RFLP) bound to gene of interest are identified and these regions are then used as a probe for searching in BAC libraries during identification of orthologous regions of large genomes (e.g. barley: 5 Gbp, or wheat: 16 Gbp)

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Genomic Colinearity



Feuillet and Keller, 2002

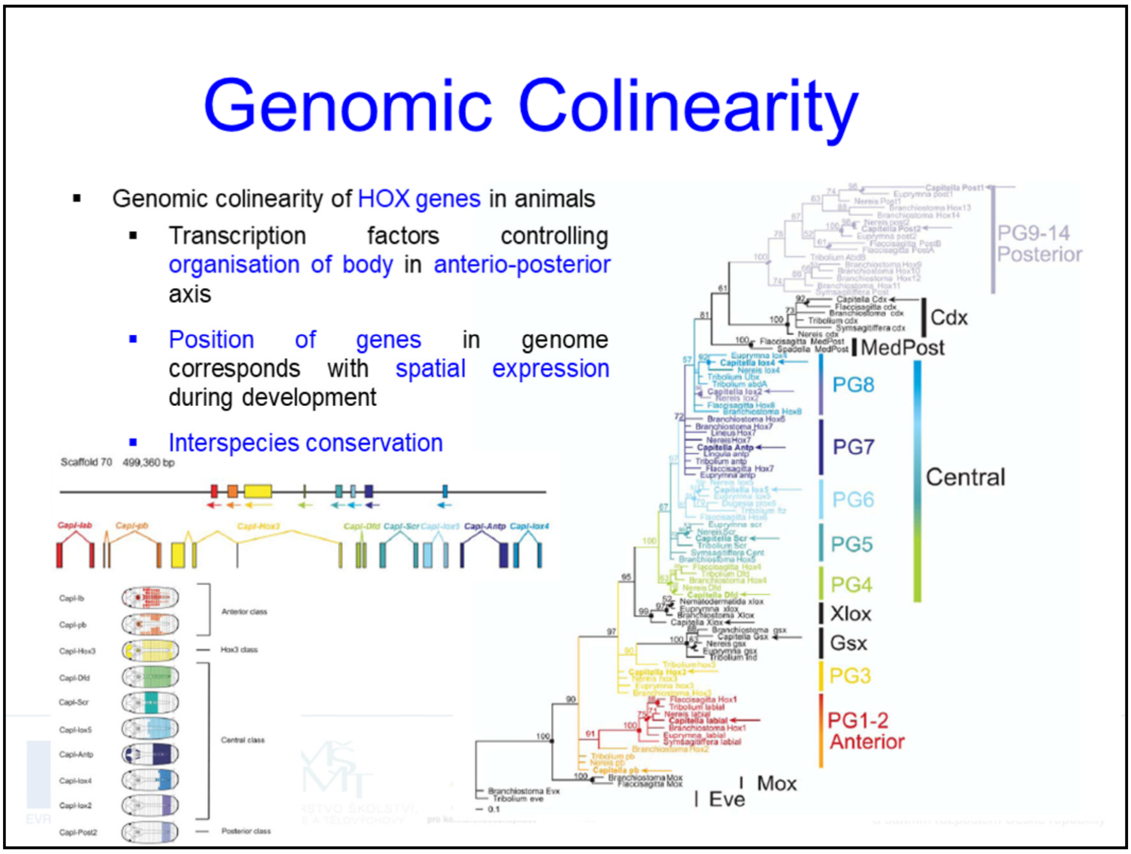


Genomic Colinearity

- Can be mostly used for the species of grass (e.g. using related genes of species of barely, wheat, rice, maize)
- Small genome reorganizations (deletions, duplications, inversions, translocations smaller than a few cM) are then detected by detailed sequentional comparative analysis
- During evolution there's occured some divergencies in related species, mostly in non-coding regions (invasion of retrotransposons etc.)

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

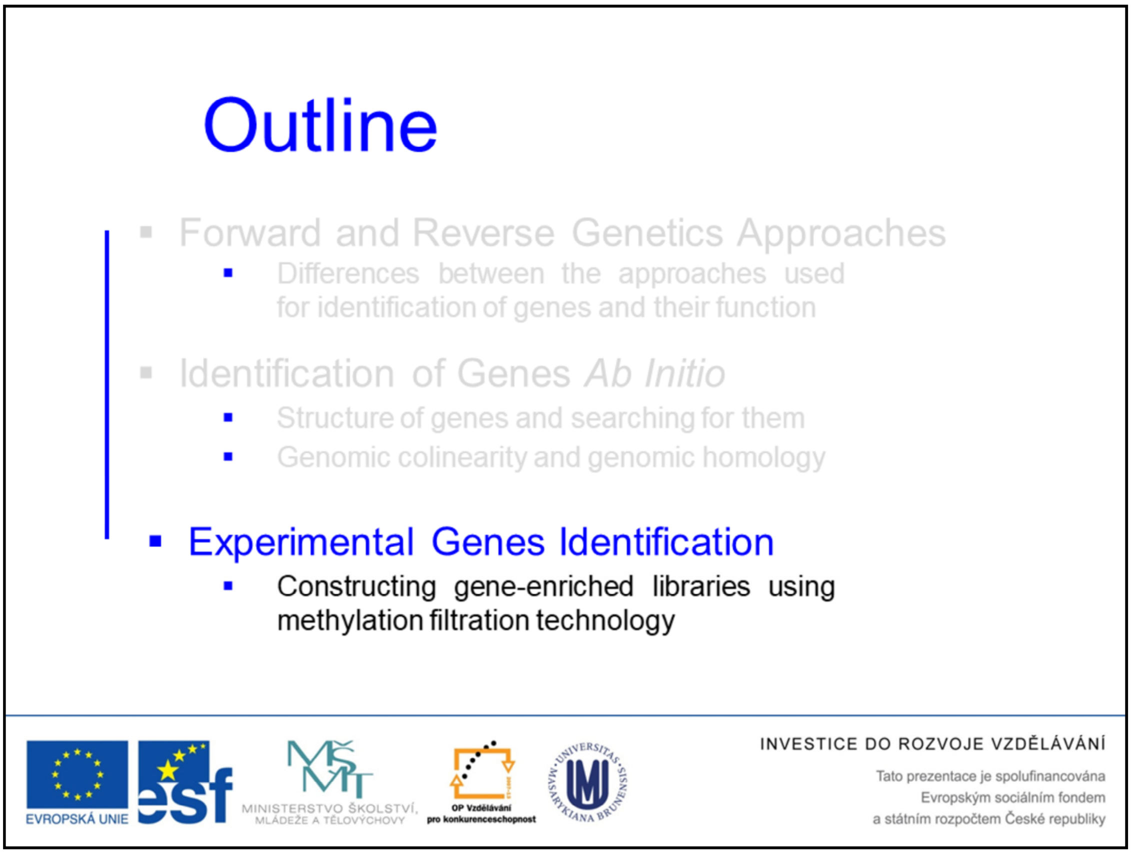


Genomic Colinearity

- Genomic colinearity of HOX genes in animals
  - Transcription factors controlling organisation of body in anterio-posterior axis
  - Position of genes in genome corresponds with spatial expression during development
  - Interspecies conservation

Genomic organization of the Capitella sp. I Hox cluster. A total of 11 Capitella sp. I Hox genes are distributed among three scaffolds. Black lines depict two scaffolds, which contain 10 of the Capitella sp. I Hox genes. The eleventh gene, CapI-Post1, is located on a separate scaffold surrounded by ORFs of non-Hox genes (unpublished data). No predicted ORFs were identified between adjacent linked Hox genes. Transcription units are shown as boxes denoting exons, connected by lines that denote introns. Transcription orientation is denoted by arrows beneath each box. Color coding is the same as that used in on the right-hand side for each ortholog.

The phylogenic tree on the right-hand side shows that the order of the genes on the chromozome is retained in several species (genome colinearity).

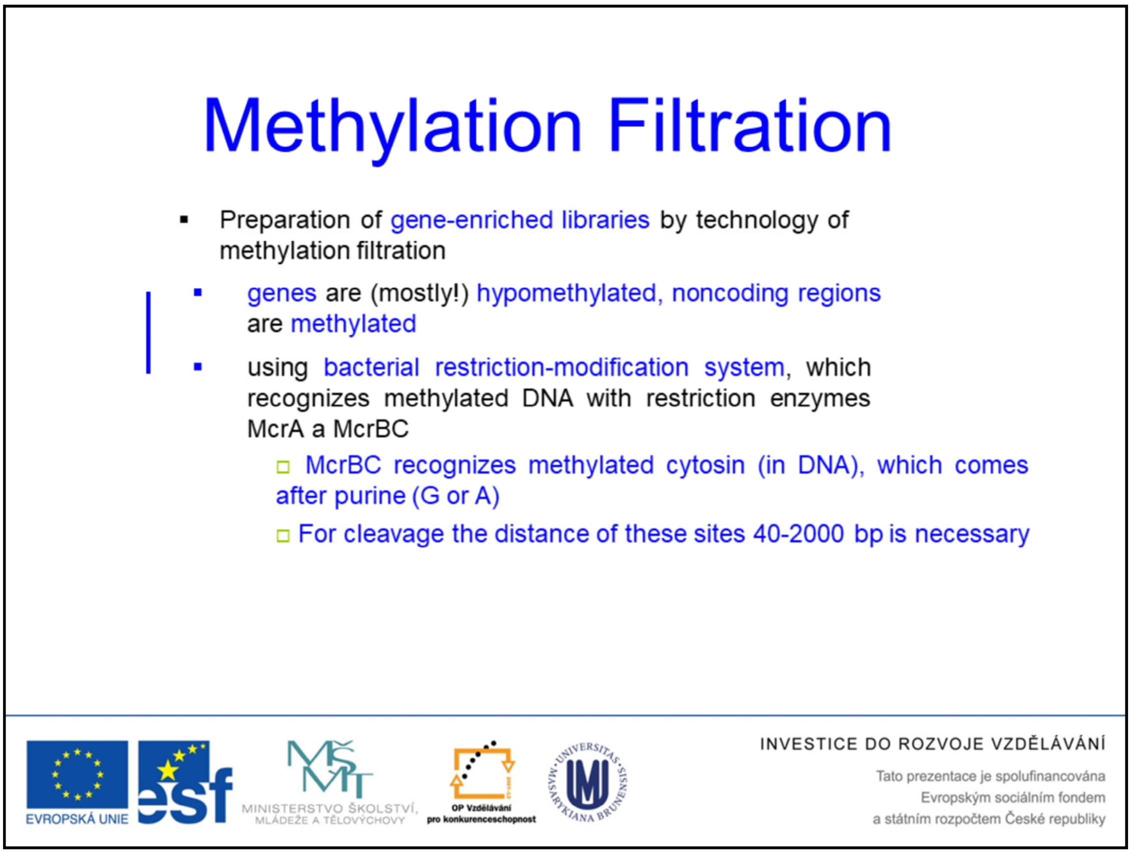


Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them
  - Genomic colinearity and genomic homology
- Experimental Genes Identification
  - Constructing gene-enriched libraries using methylation filtration technology

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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



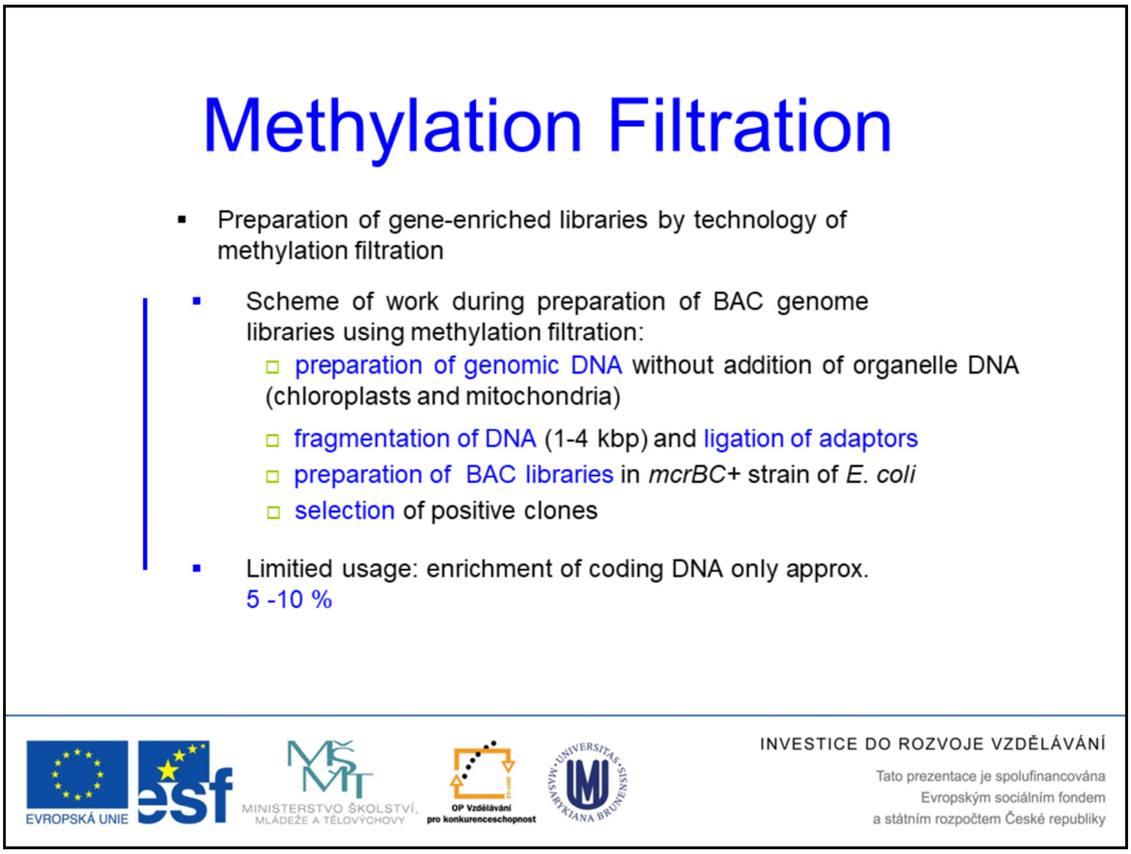
Methylation Filtration

- Preparation of gene-enriched libraries by technology of methylation filtration
  - genes are (mostly!) hypomethylated, noncoding regions are methylated
  - using bacterial restriction-modification system, which recognizes methylated DNA with restriction enzymes McrA a McrBC
    - □ McrBC recognizes methylated cytosin (in DNA), which comes after purine (G or A)
    - □ For cleavage the distance of these sites 40-2000 bp is necessary

EVROPSKÁ UNIE
esf
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

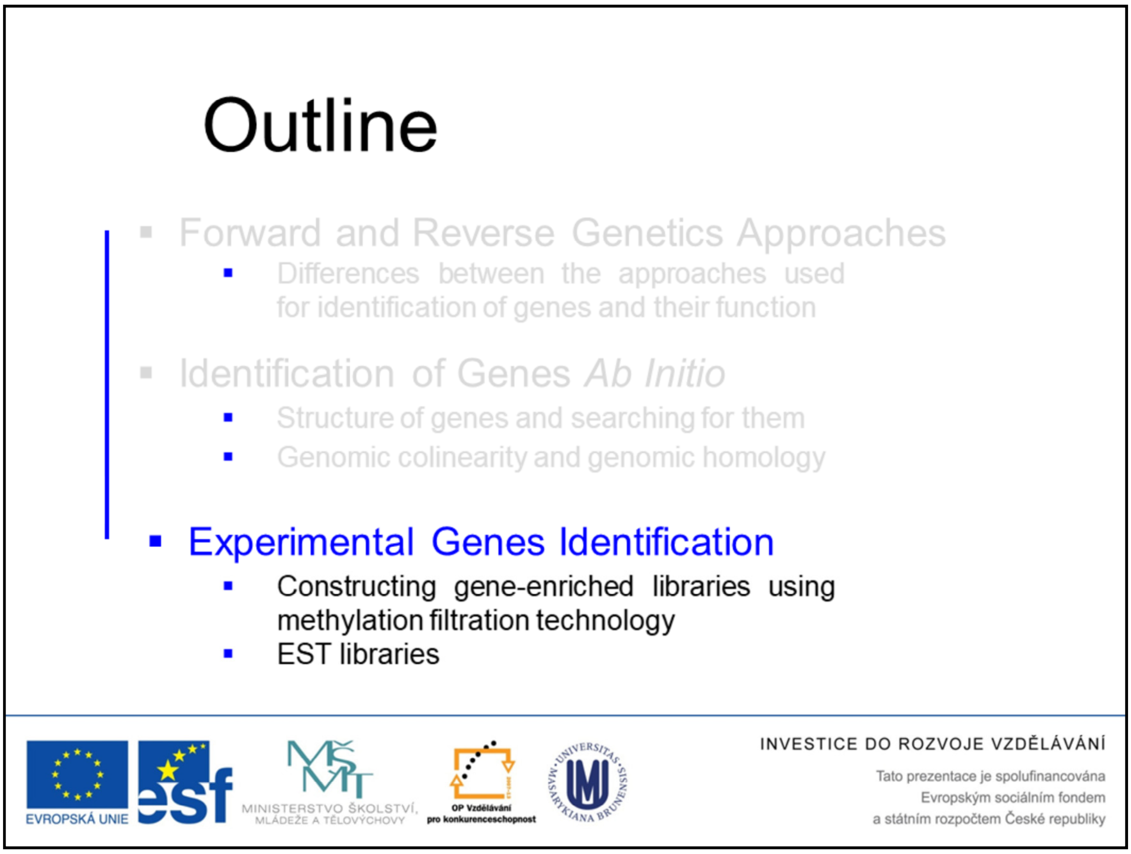
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Methylation Filtration

- Preparation of gene-enriched libraries by technology of methylation filtration
  - Scheme of work during preparation of BAC genome libraries using methylation filtration:
    - □ preparation of genomic DNA without addition of organelle DNA (chloroplasts and mitochondria)
    - □ fragmentation of DNA (1-4 kbp) and ligation of adaptors
    - □ preparation of BAC libraries in *mcrBC*+ strain of *E. coli*
    - □ selection of positive clones
  - Limitied usage: enrichment of coding DNA only approx. 5 -10 %

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky



Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them
  - Genomic colinearity and genomic homology
- Experimental Genes Identification
  - Constructing gene-enriched libraries using methylation filtration technology
  - EST libraries

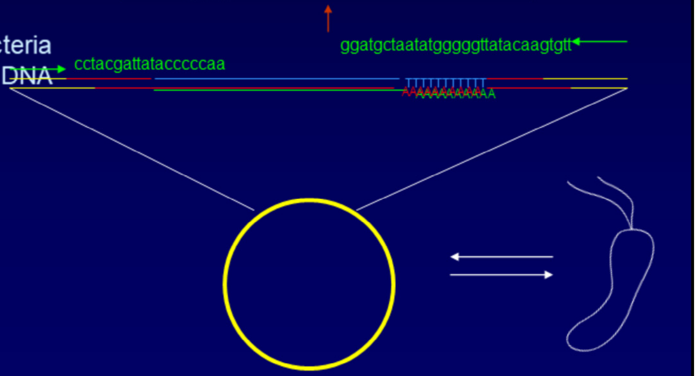
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky

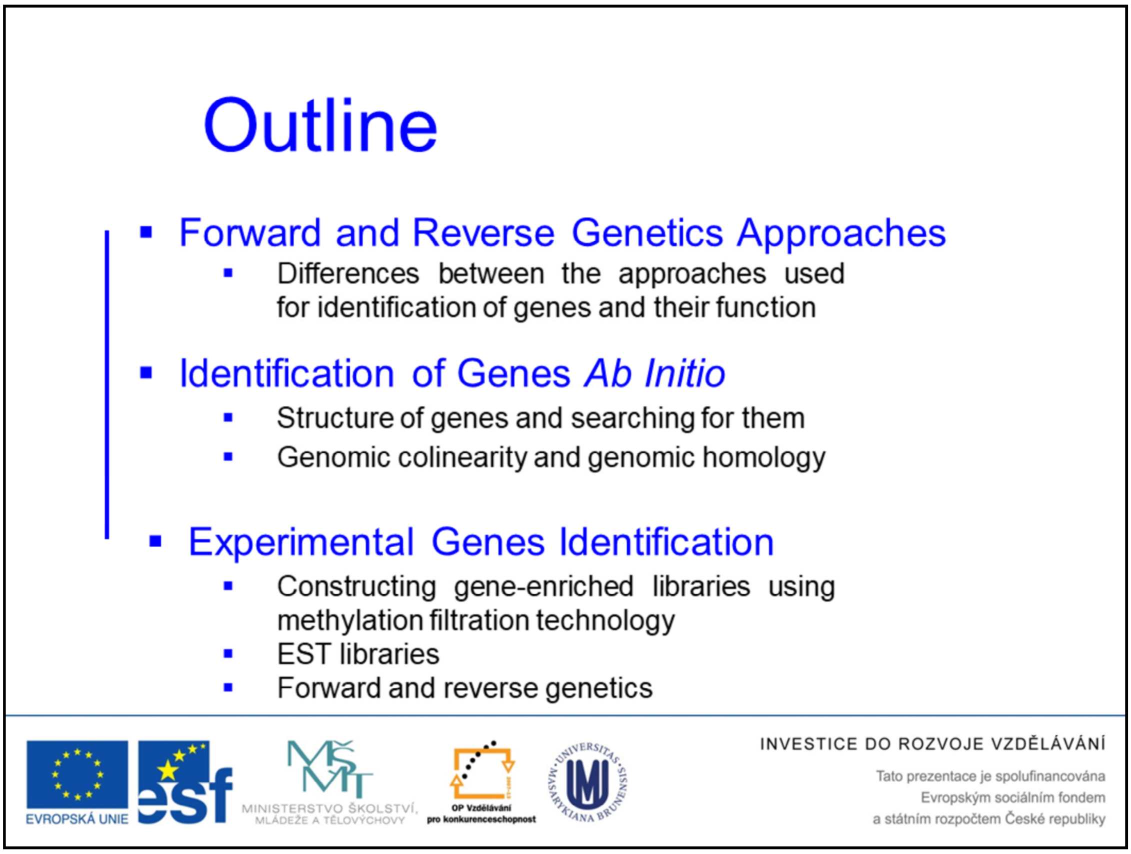
EST Libraries

- Preparation of EST libraries
  - Isolation of mRNA
  - Reverse transcription
  - Ligation of linkers and synthesis of second cDNA strand
  - Cloning into suitable bacterial vector
  - Transformation into bacteria and isolation of DNA (amplification of DNA)
  - Sequencing using primers specific for used plasmid
  - Saving the results of sequencing into public database

NC_002377. Agrobacterium tum. [gi:10955016]

[illegible]





Outline

- Forward and Reverse Genetics Approaches
  - Differences between the approaches used for identification of genes and their function
- Identification of Genes *Ab Initio*
  - Structure of genes and searching for them
  - Genomic colinearity and genomic homology
- Experimental Genes Identification
  - Constructing gene-enriched libraries using methylation filtration technology
  - EST libraries
  - Forward and reverse genetics

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ
Tato prezentace je spolufinancována
Evropským sociálním fondem
a státním rozpočtem České republiky

Discussion

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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