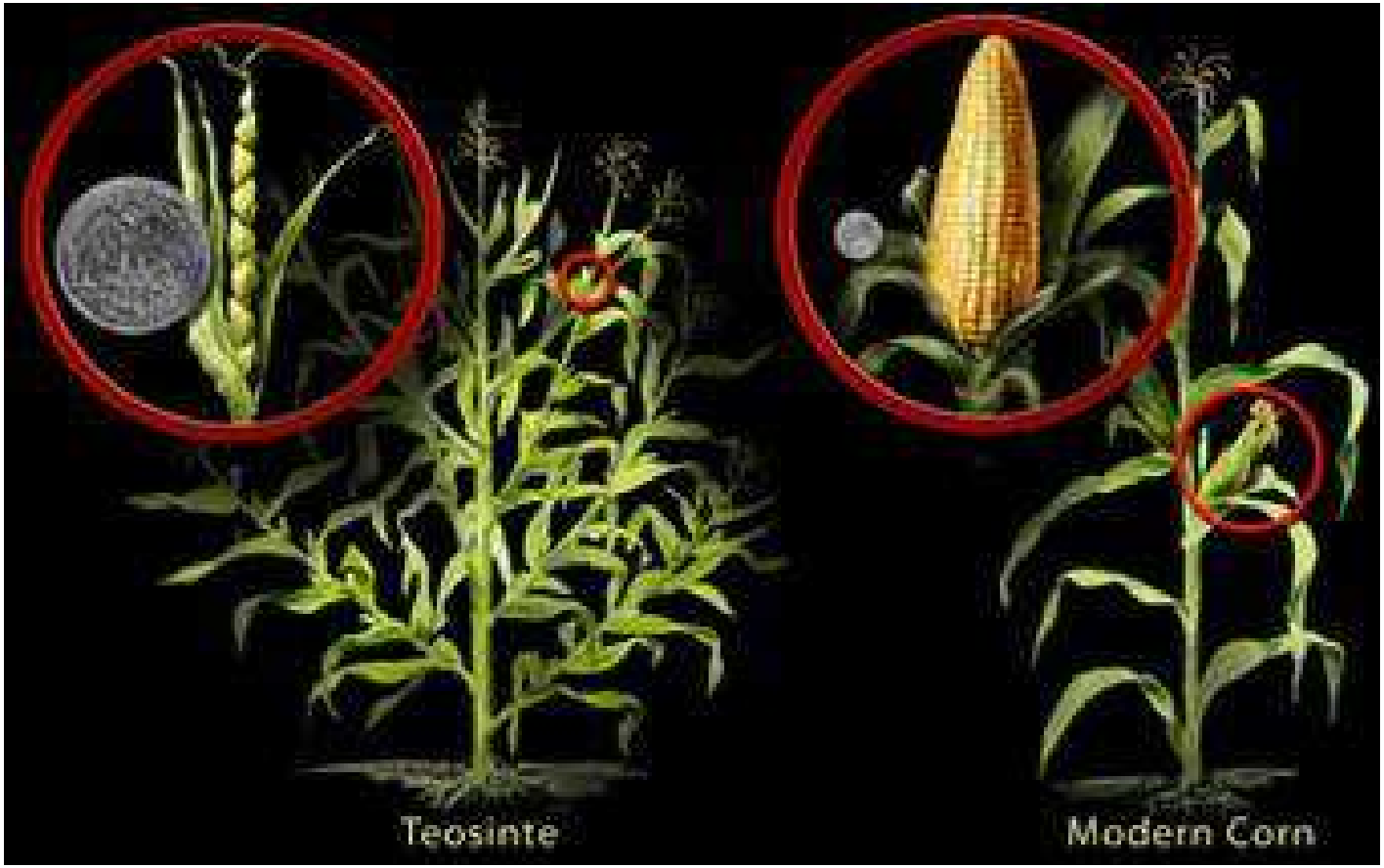
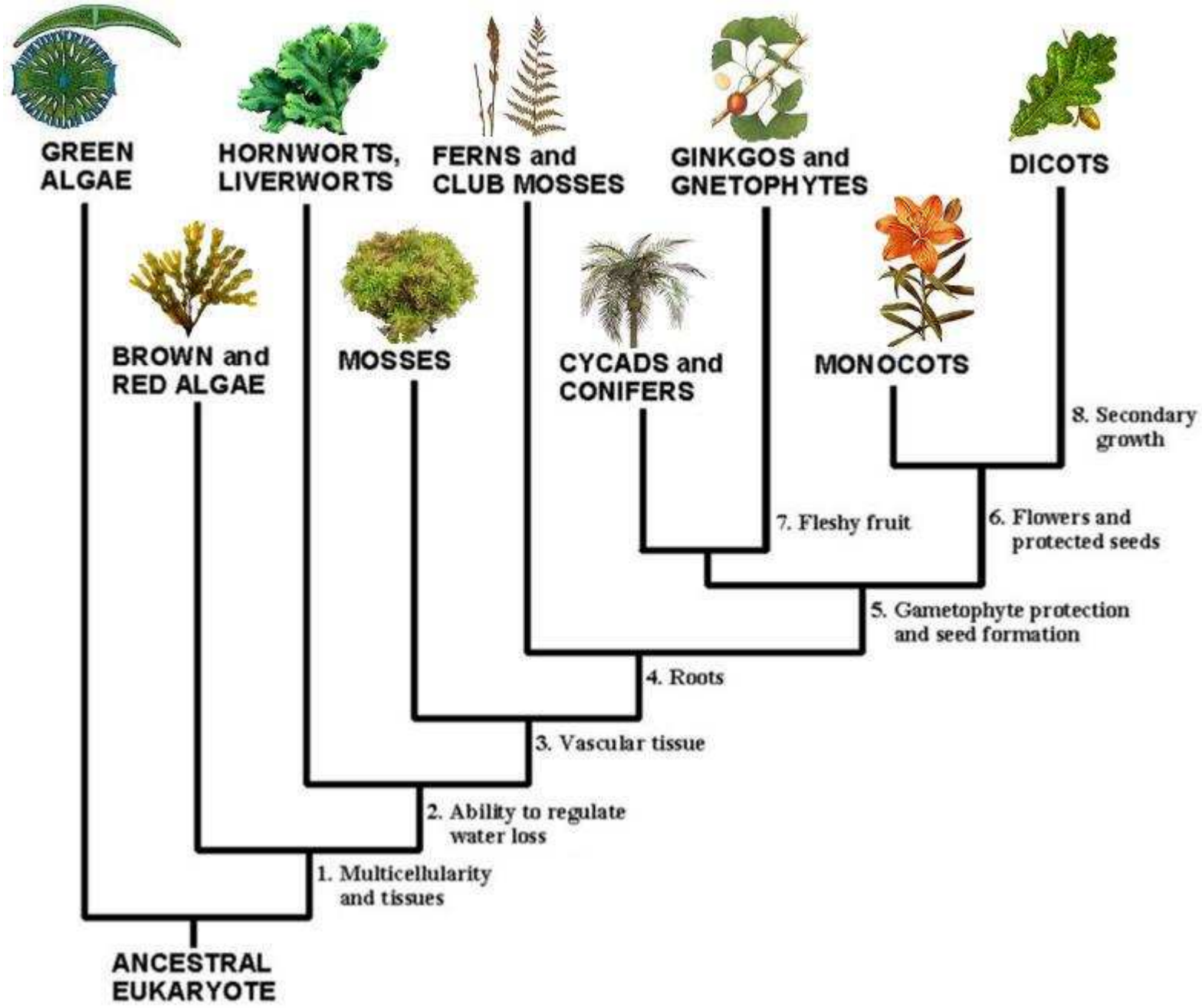
STUDIO DEI GENOMI VEGETALI

- **Identificazione di geni importanti per caratteri agronomici (produttività, resistenza a stress, proprietà nutrizionali)**

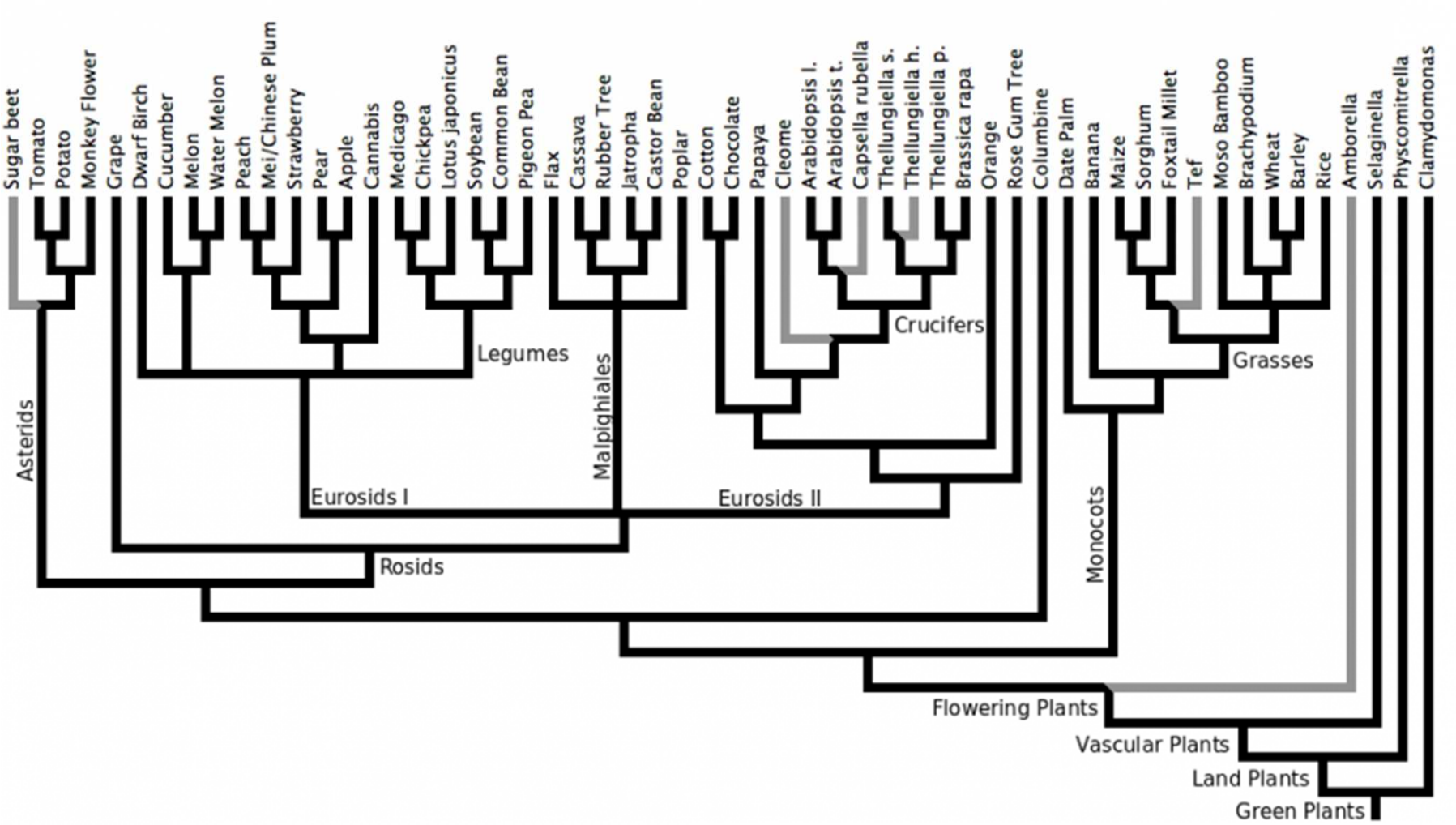
- **Comprensione dell'evoluzione delle piante**

Addomesticamento delle specie vegetali




GREEN ALGAE
HORNWORTS, LIVERWORTS
FERNS and CLUB MOSSES
GINKGOS and GNETOPHYTES
DICOTS
BROWN and RED ALGAE
MOSSES
CYCADS and CONIFERS
MONOCOTS
8. Secondary growth
7. Fleshy fruit
6. Flowers and protected seeds
5. Gametophyte protection and seed formation
4. Roots
3. Vascular tissue
2. Ability to regulate water loss
1. Multicellularity and tissues
ANCESTRAL EUKARYOTE

http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes



Quali specie sequenziare?

- Impatto economico, sociale e scientifico
- Distanza filogenetica da altre specie sequenziate (-> nuove informazioni)
- Informazioni disponibili (mappe genetiche e fisiche)
- Capacità di persuasione dei ricercatori

Arabidopsis thaliana

Dicotiledone (*Brassicaceae*)

Piccolo genoma dipolide (C1 = 125 Mbp)

Trasformabile facilmente

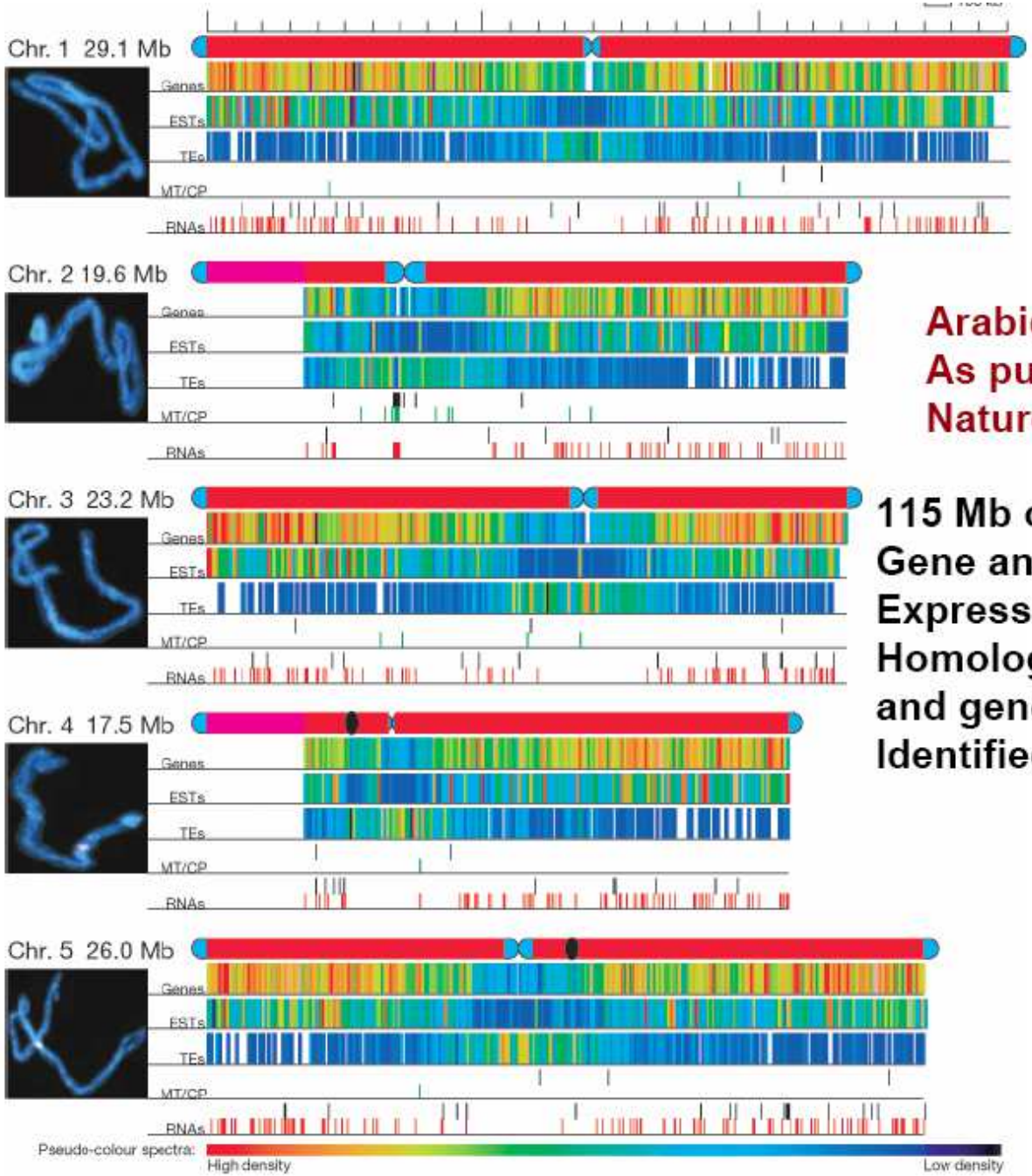
5 cromosomi

Piccole dimensioni

Ciclo vitale breve (2 mesi)

Primo genoma sequenziato

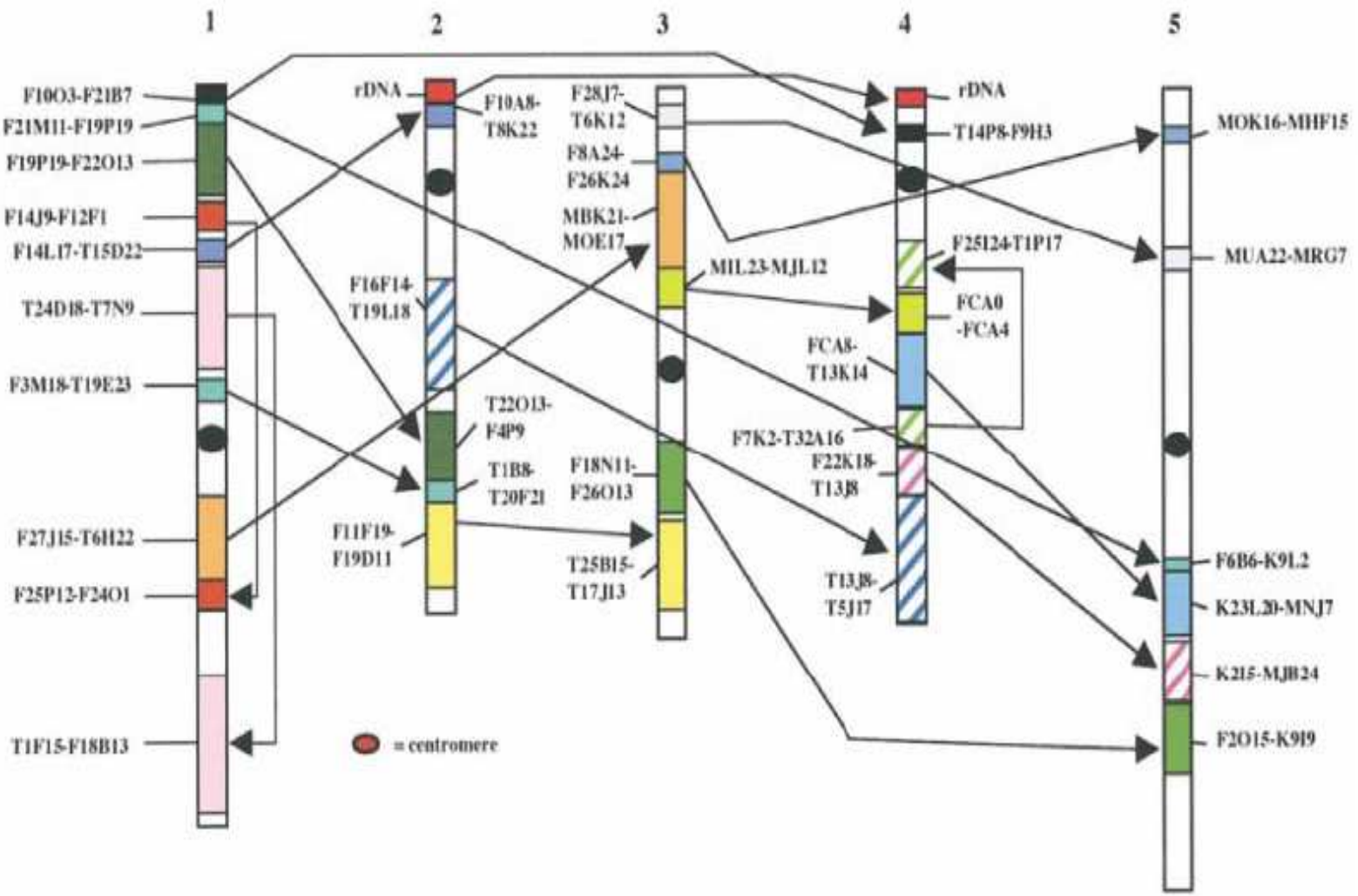
The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408 (6814), 796-815 DOI: 10.1038/35048692



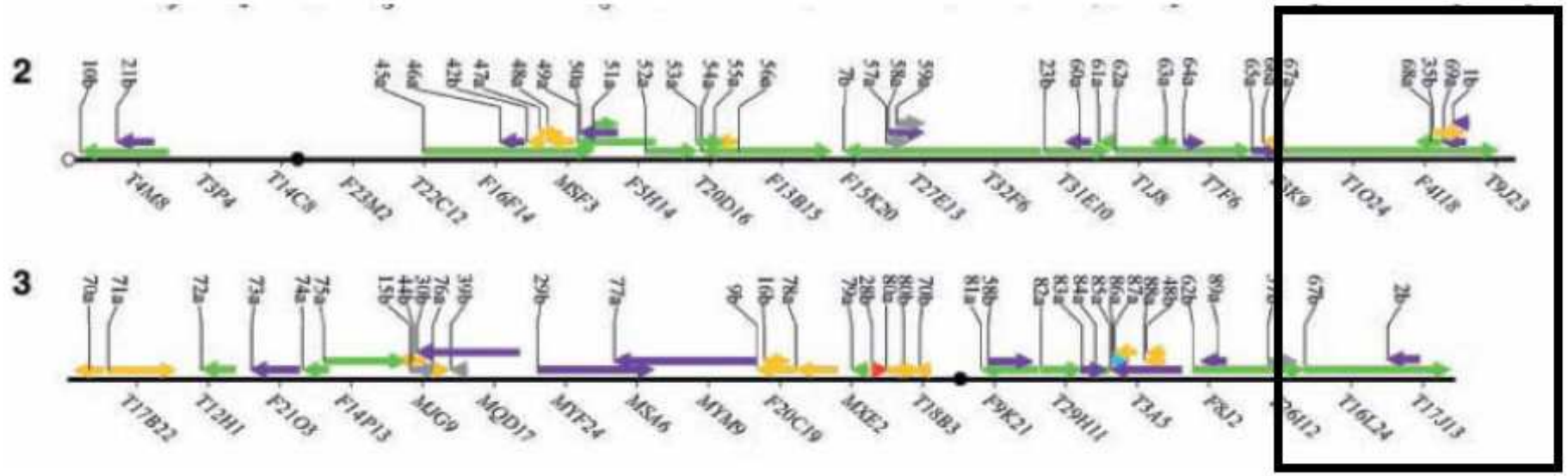
Arabidopsis genome sequence.
As published in 2000.
Nature 408, 796.

115 Mb of 125 Mb genome.
Gene annotation using
Expressed sequence tags (ESTs)
Homology with cloned plant genes and genes of other organisms
Identified 25,500 genes.

Large segments of the Arabidopsis genome are duplicated



Genetic redundancy can exist between genes in duplicated blocks



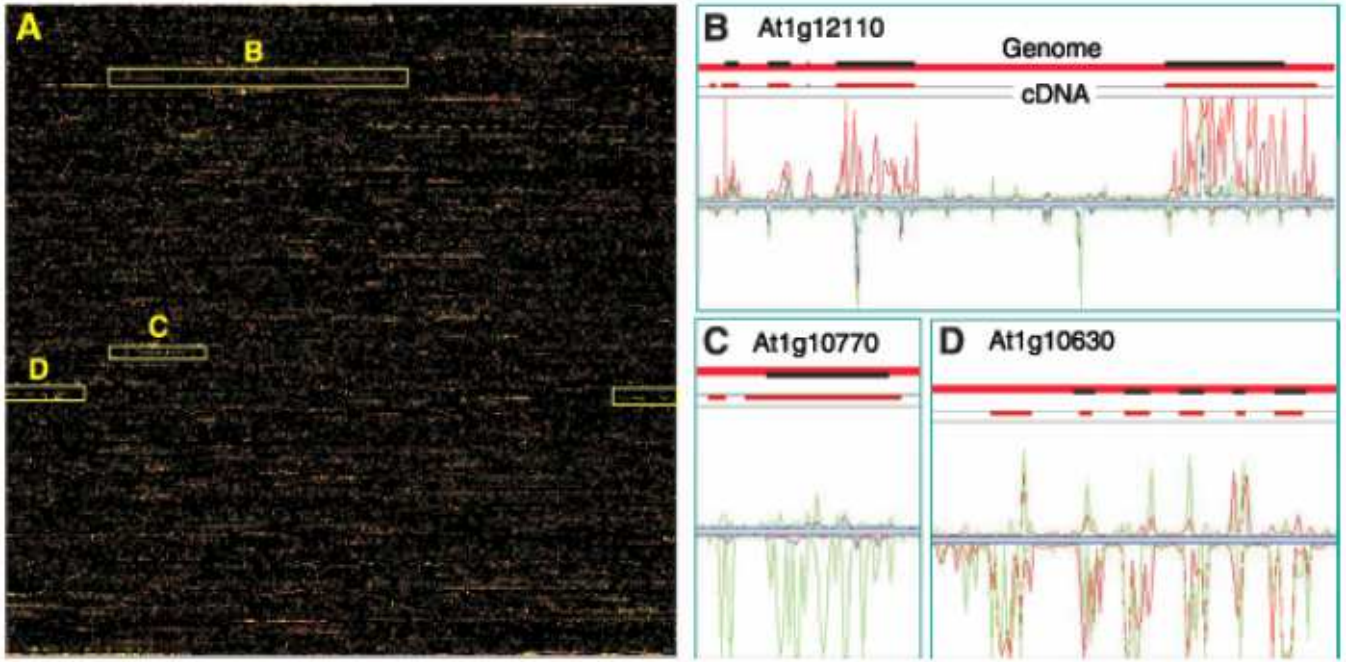
A duplicated block of genes exists on chromsomes 2 and 3.

One of the duplicated genes encodes a MADS box transcription factor, and the proteins encoded by the two genes are 87% identical at the amino acid level.

Improved annotation of the Arabidopsis genome sequence

Initial sequence analysis relied heavily on expressed sequence tags (ESTs) and gene prediction programmes. Interpretation of genome sequence constantly improved.

Isolation of full-length cDNAs and use of whole genome oligonucleotide tiling arrays greatly improved the annotation. (32% of predicted gene models incorrect).



Arabidopsis genome represented on 12 oligonucleotide arrays.
Each array 834,000 25-mer oligonucleotides.
Is hybridized with cDNA made from 4 RNA populations; seedlings, roots, flowers, cultured cells. Sequences that hybridize are mapped directly onto genome sequence.

Enzymes involved in secondary metabolism

Arabidopsis genome contains many classes of enzymes involved in secondary metabolism that are required for the synthesis of specialized compounds.

An example, is the family of genes encoding the Cytochrome P450 monoxygenase enzymes.

Mammals, C.elegans, Drosophila – 80 – 105 genes.

Arabidopsis – 246 genes.

In plants these enzymes are required for the synthesis of compounds such as growth regulators (gibberellic acid, Brassinosteroid), carotenoids (protect cell from oxidative damage) and phenylpropanoids that are present in plant cell walls.

Transcription factors

Arabidopsis contains around 1500 genes encoding transcription factors (aprox. 5%)

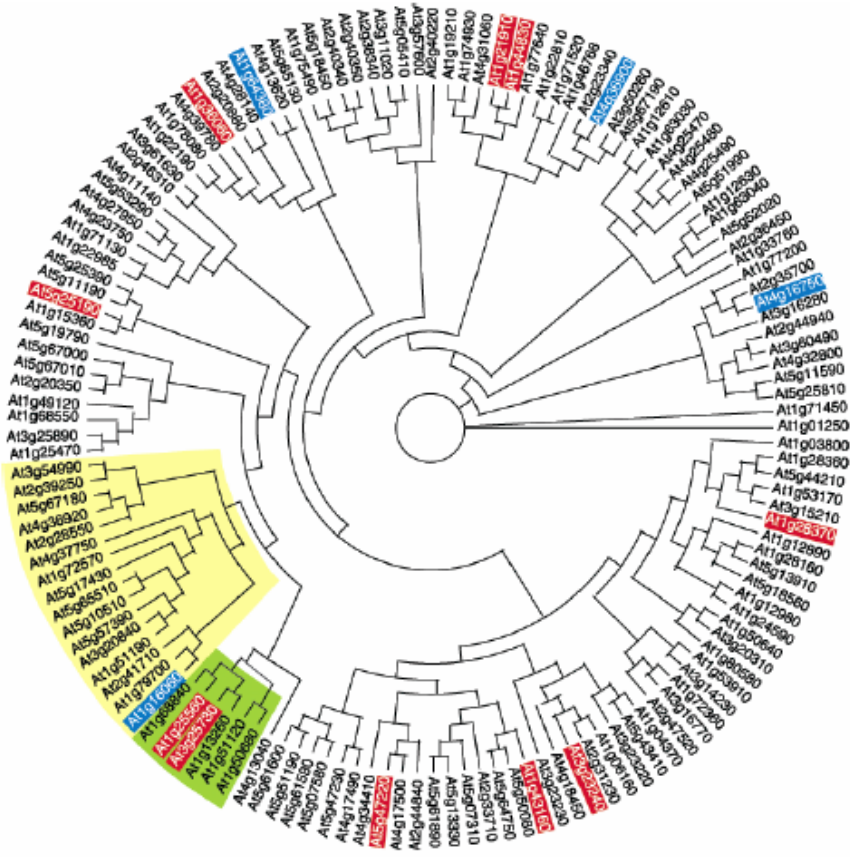
Drosophila contains around 640 genes encoding transcription factors, around 4.5%.

Many important animal transcription factor families are absent in plants, such as nuclear steroid receptors, NHR zinc finger proteins (252 in C. Elegans) and Fork head transcription factors (18 in Drosophila, 15 in C.elegans).

Each eukaryotic lineage has its own set of transcription factor families.

APETALA 2-like transcription factors are unique to plants

This is one of several families of transcrition factor that are only found in plants.



141 AP2-domain TFs in Arabidopsis

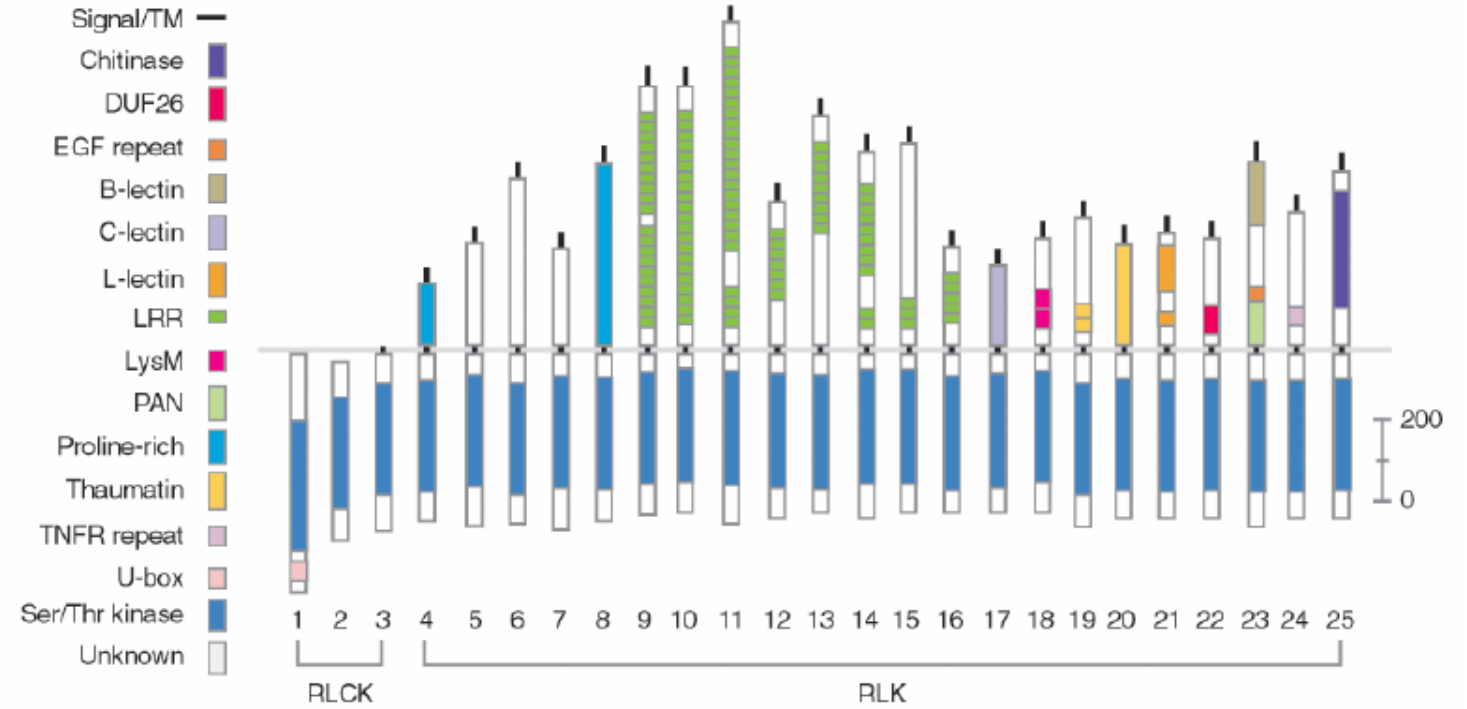
-Some related to APETALA2, Involved in flower development (yellow)

-Some involved in ethylene response and their transcription is changed by ethylene (red/blue)

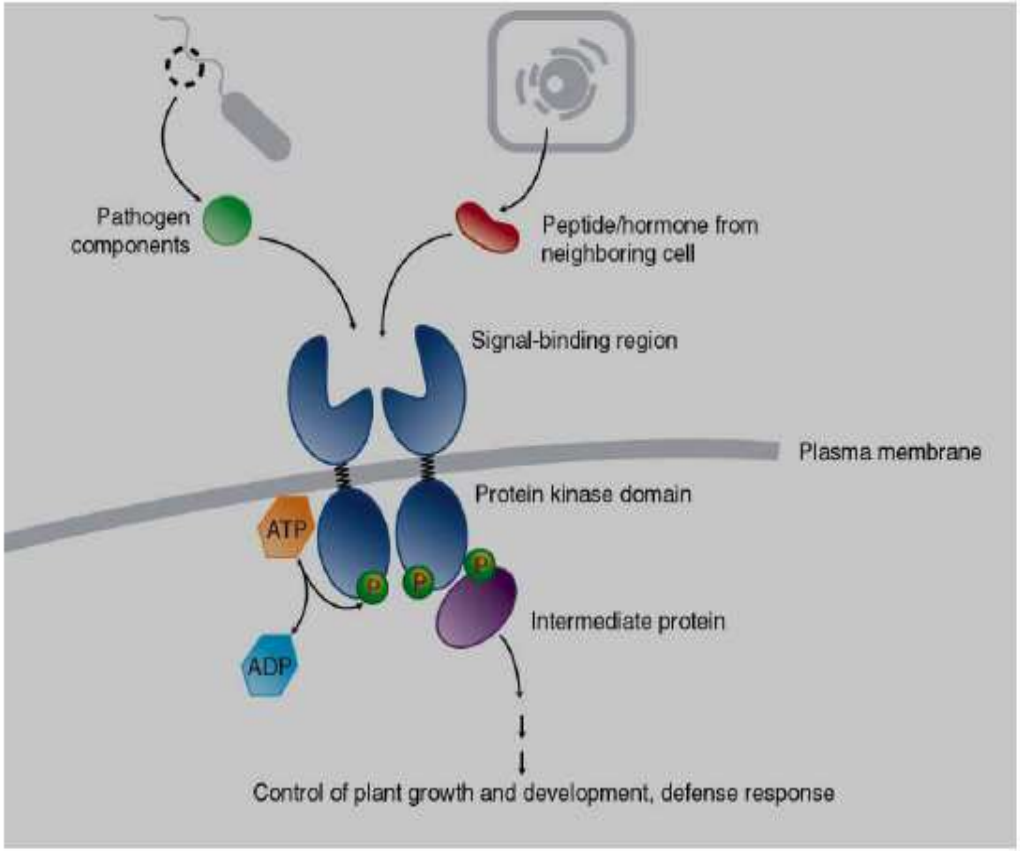
-Others are involved in stress responses to drought

Receptor-like kinases

600 Arabidopsis genes encode receptor-like kinases predicted to be located in the membrane.
These are similar in domain organization to animal receptor tyrosine kinases, such as epidermal growth factor.
However, are predicted to be serine/threonine kinases, and have divergent ligand binding domains.



Some Receptor Like Kinases have important functions, but for most their function is unknown



Many RLKs are of unknown function. However, some have defined roles:

Brassinosteroid receptor
Clavata 1
Resistance to pathogens

Other plant-specific processes

Hundreds of genes involved in photosynthesis

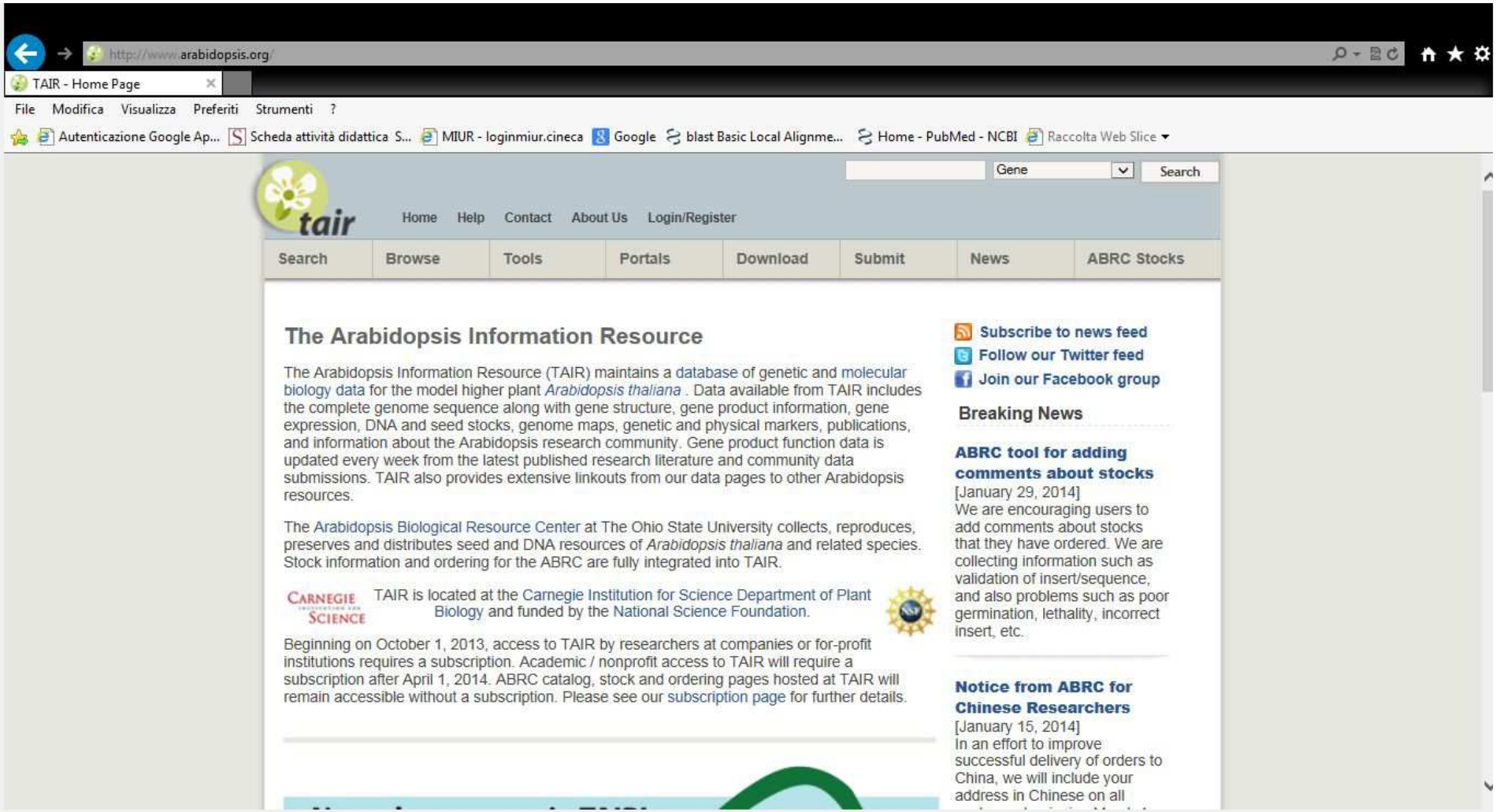
- **light harvesting**
- **chlorophyll biosynthesis**
- **carbon dioxide fixation**
- **energy generating photosystems**

Transporters

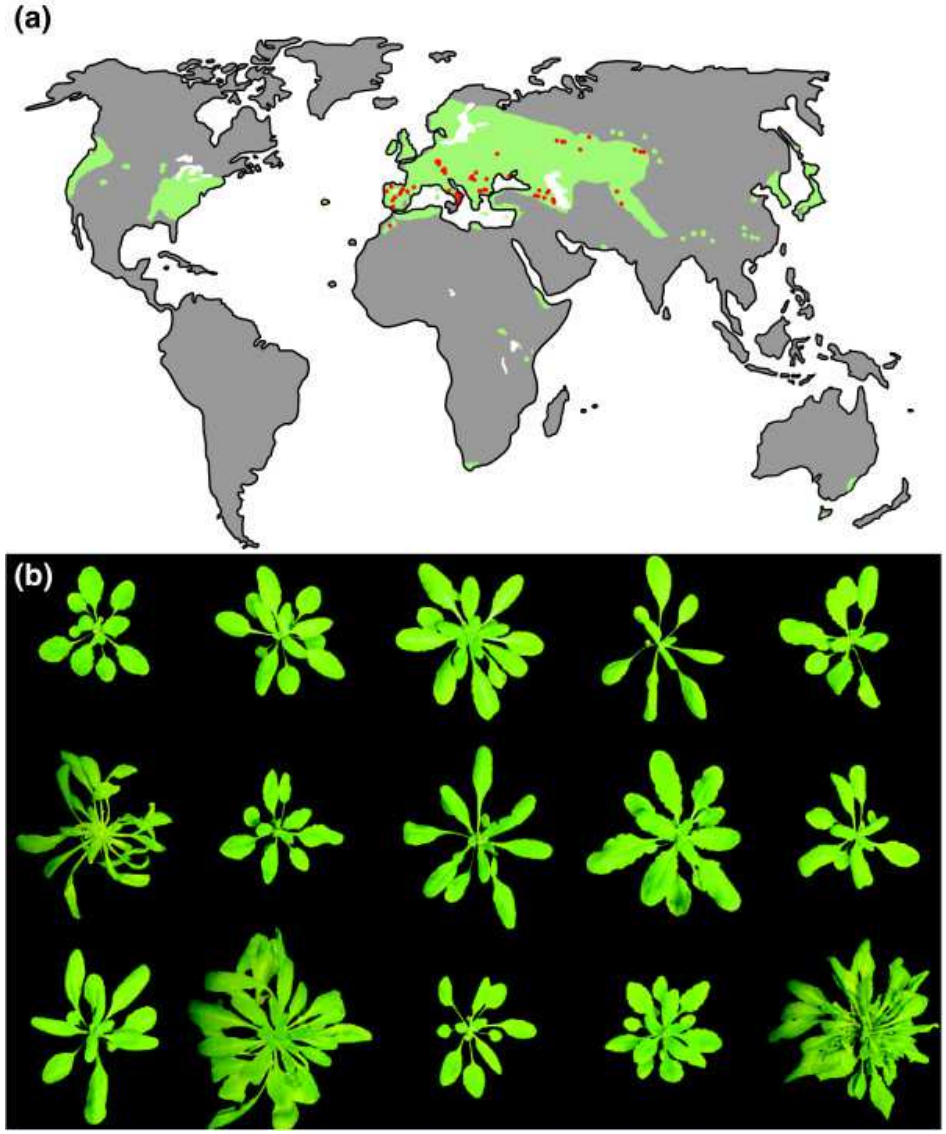
- **plants mainly use proton-type ATPases, whereas in animals transport is usely coupled with sodium ions via sodium-type ATPases.**

No major histocompatibility complex, however 100s of genes Encoding nucleotide binding site leucine rich repeat proteins Involved in pathogen resistance.

The Arabidopsis Information Resource (TAIR)
http://www.arabidopsis.org/



1001 Genomes : A Catalog of Arabidopsis thaliana Genetic Variation
http://1001genomes.org/accessions.html

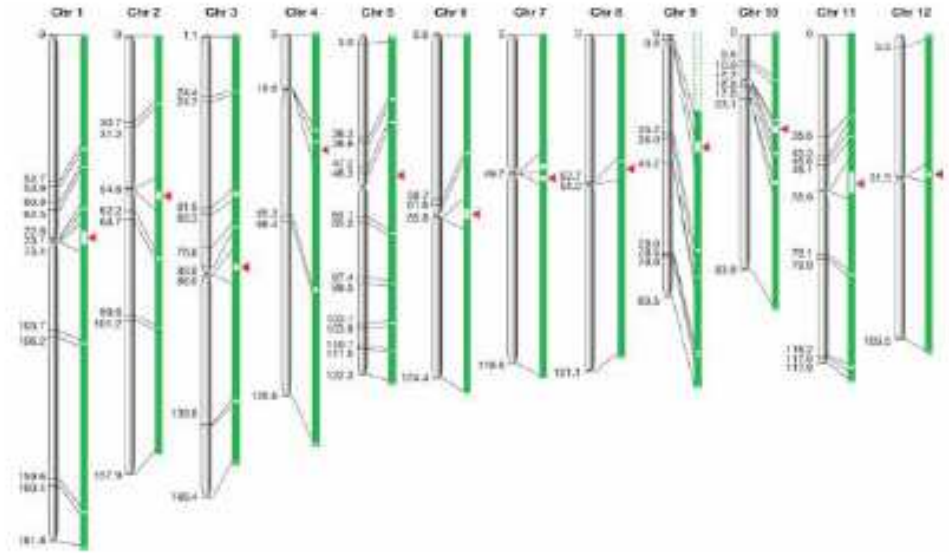


818 accessioni sequenziate e rilasciate al 13-3-2014

Weigel and Mott *Genome Biology* 2009, 10:107
doi:10.1186/gb-2009-10-5-107

RISO (*Oryza sativa*)

The rice genome



389 Mbp, 12 Chromosomes
Finished sequence (95% coverage)

37,544 / 22,840 [61%] genes (FGENESH models / cDNA-supported)
71% / 89% have homologs in Arabidopsis
90% of Arabidopsis genes have homologs in rice

2,859 genes not present in Arabidopsis

80,127 polymorphic loci *japonica / indica*

"With a large number of proteins of unknown function, the most interesting differences between the genome content of [Arabidopsis and rice] remain to be discovered."

International Rice Genome Sequencing Project, Nature (2005)

GENOMICA COMPARATIVA

- Analisi e confronto di genomi di specie diverse
- Fornisce informazioni sull'evoluzione delle specie e sulla funzione di geni e sequenze non codificanti
- Es.: funzione di un gene dedotta dallo studio di geni ortologhi in specie modello

GENOMICA COMPARATIVA
Cosa si analizza?

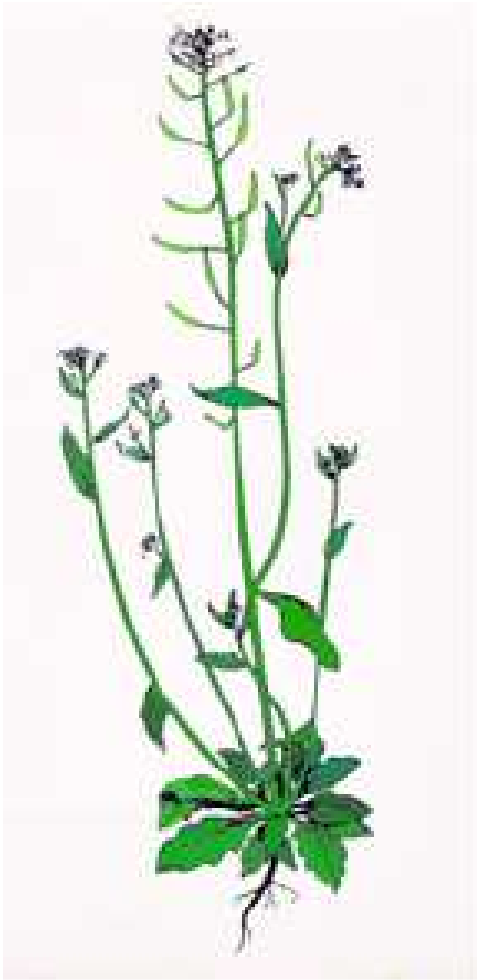
- Similarità di sequenza
- Localizzazione cromosomica dei geni
- Lunghezza e numero esoni
- Quantità di DNA non codificante
- Conservazione di regioni cromosomiche

Ostacoli al sequenziamento di specie coltivate

- **Dimensioni**
- **DNA ripetitivo**
- **Poliploidia**

Dimensioni del genoma

Arabidopsis: 125 Mb

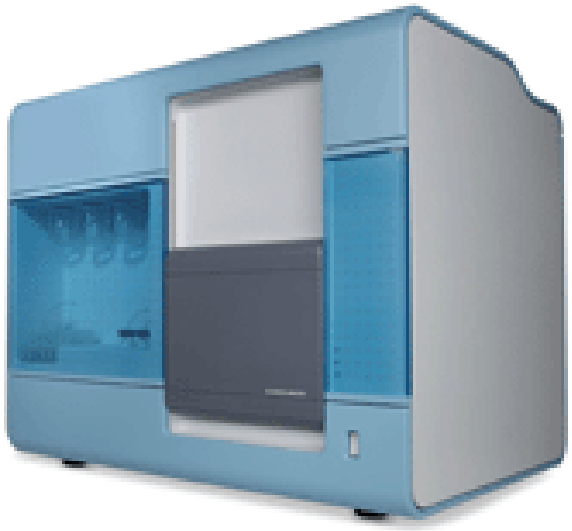


***Fritillaria assyriaca*: 125 Gb!**

ILLUMINA GENOME ANALYZER

Permette il sequenziamento in parallelo di un numero massiccio di frammenti genomici

-> 1 milione di basi sequenziate per volta!



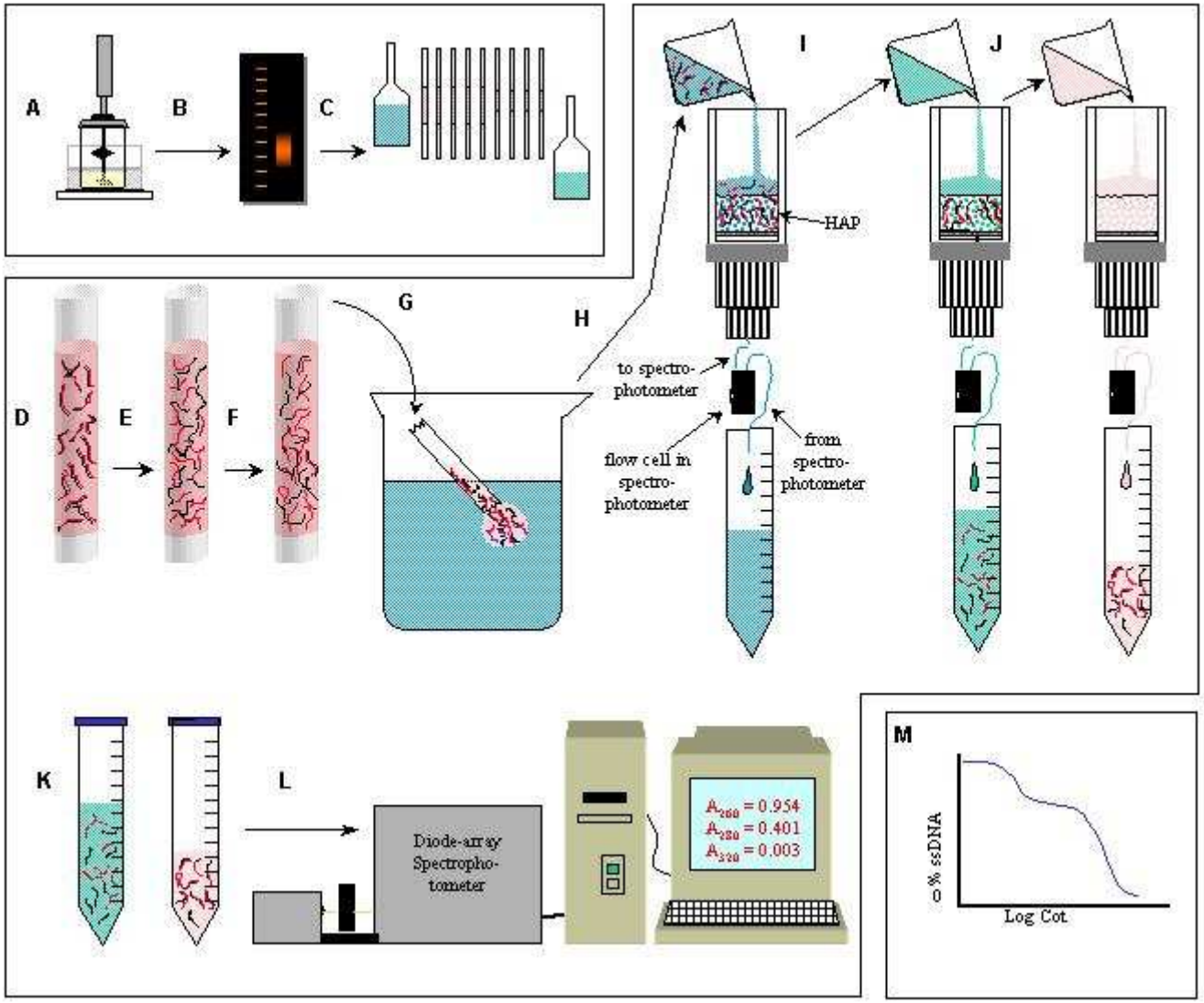
DNA ripetitivo

- Responsabile per gran parte della variabilità nelle dimensioni del genoma vegetale
- Complica l'assemblamento delle sequenze
- Sequenze non-ridondanti nel genoma: da 13% (cipolla) a 77% (pomodoro)

DNA ripetitivo

N.B.: le piante hanno più DNA ripetitivo degli animali, e copie individuali possono avere meno mutazioni per distinguerle, perchè più recenti

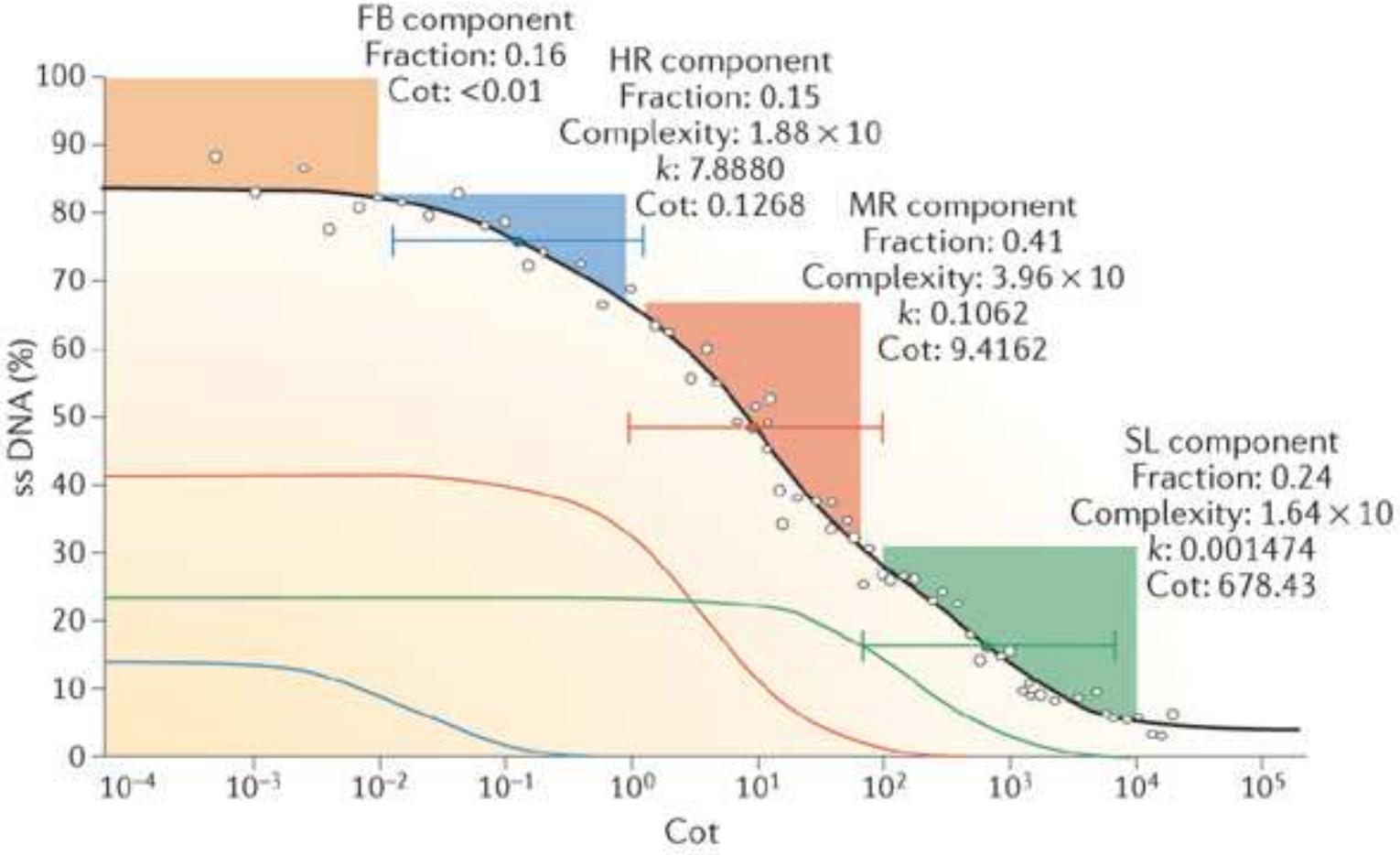
Cinetica di riassociazione



Cinetica di riassociazione

- Fornisce il valore Cot, cioè il prodotto fra la concentrazione dei nucleotidi (Co) ed il tempo di riassociazione (normalizzato per la conc. di cationi nel tampone)
- La cromatografia su colonna di idrossiapatite (che lega il dsDNA) permette di isolare la frazione di DNA che si riassocia ad un particolare valore di Cot.
- Più il DNA è ripetitivo, più basso sarà il suo valore Cot

Cinetica di riassociazione



Nature Reviews | Genetics

Clonaggio basato sul valore Cot (CBCS)

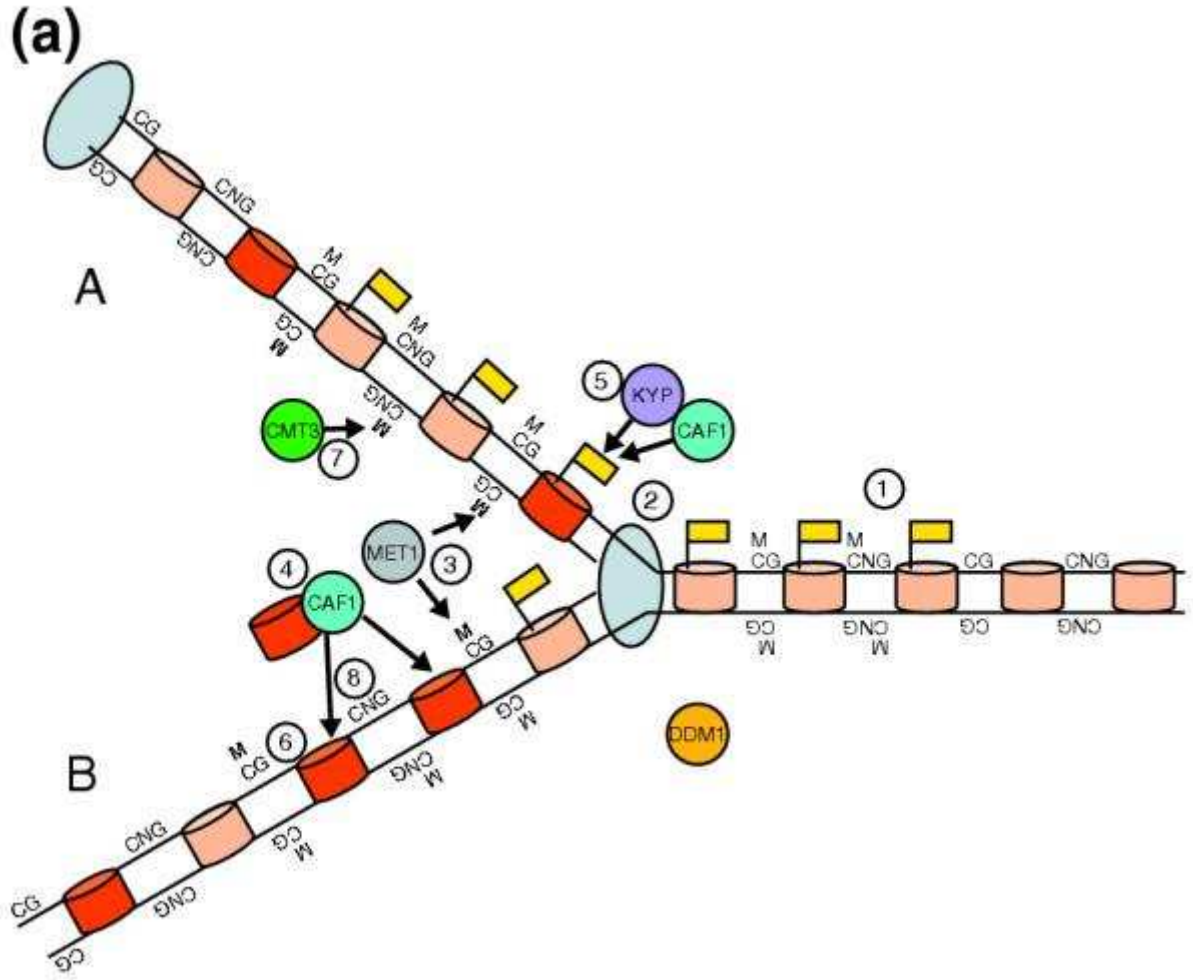
L’analisi Cot permette di isolare specificamente frazioni più o meno ripetitive

DNA meno ripetitivo viene sequenziato

-> più facile da assemblare in contigs

-> maggiore percentuale di geni

il DNA più ricco in geni è ipometilato rispetto a quello non codificante (inclusa una parte di DNA ripetitivo)



Methylation filtration (MF)

Clonaggio del DNA genomico totale in ceppi di *E. coli* che degradano il DNA metilato -> sequenziamento dei cloni e assemblaggio in contigs

Svantaggio: non sempre il DNA codificante è ipometilato (es. metilazione indotta da stress, o in colture cellulari)

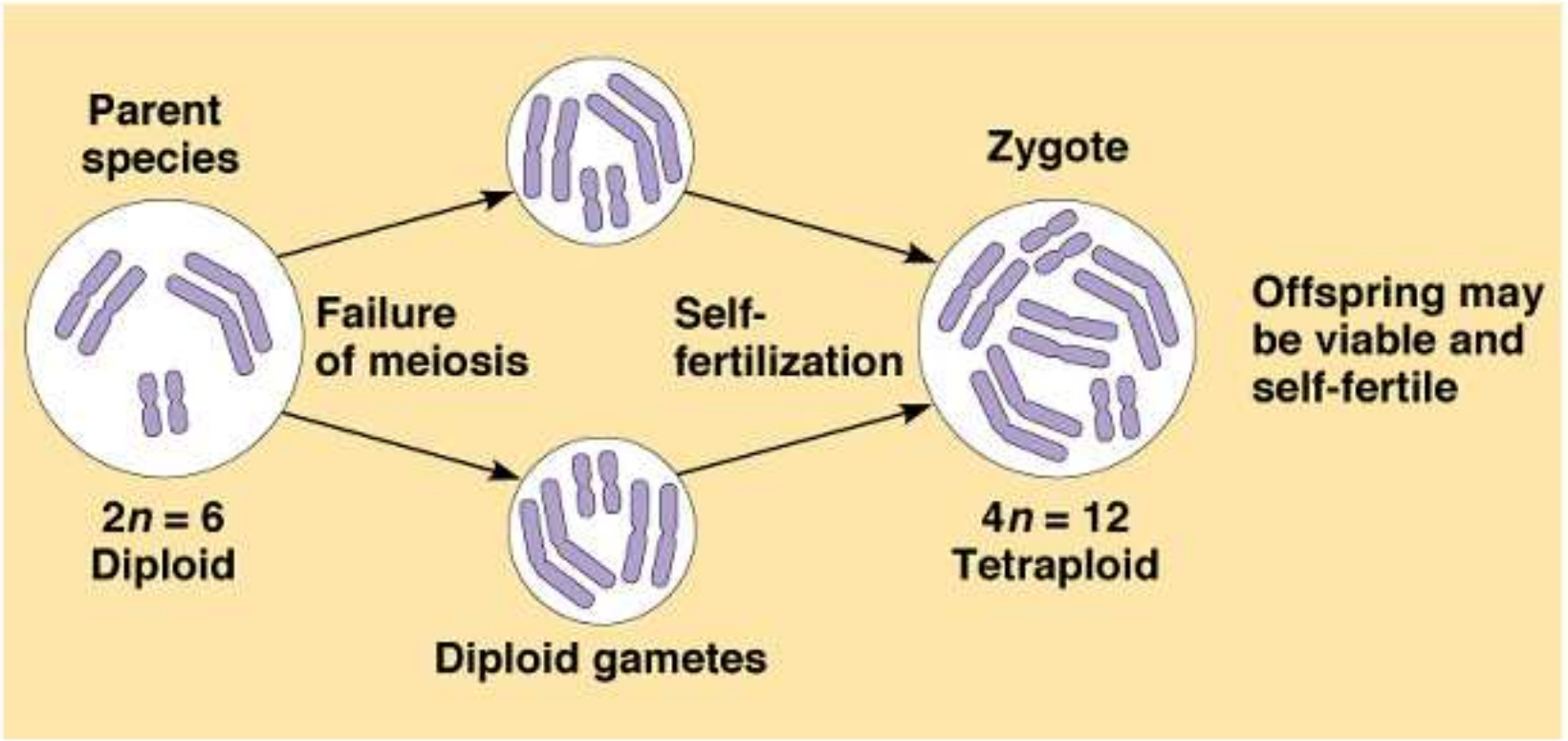
POLIPLOIDIA

- Duplicazione del genoma in una specie (**autopoliploidia**), attraverso errore meiotico (4 copie di ogni cromosoma)

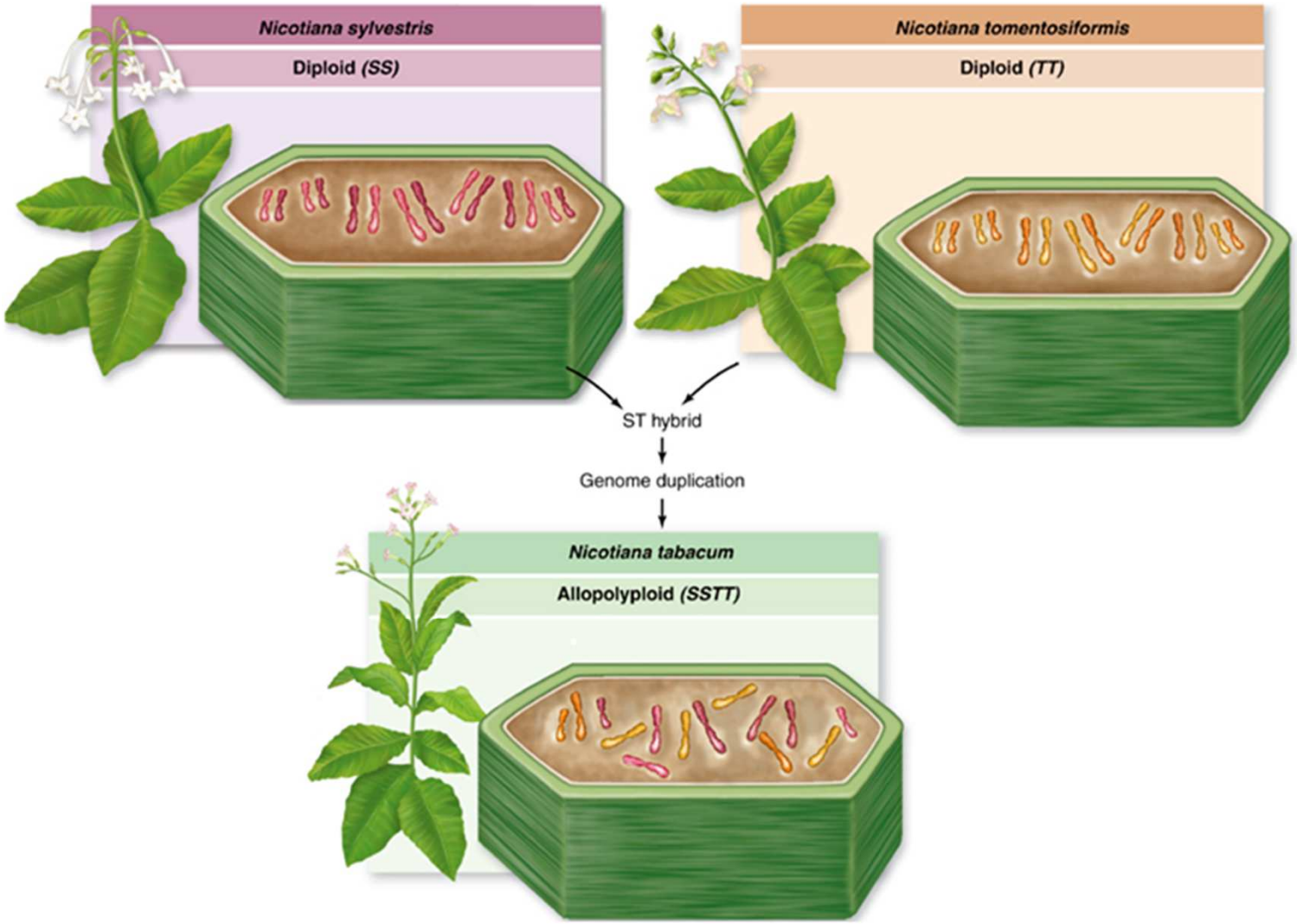
- Ibridazione di due specie diverse (**allopoliploidia**)

AUTOPOLIPLOIDIA

(es. Canna da zucchero, patata, erba medica, caffè)



ALLOPOLIPLOIDIA: TABACCO



ALLOPOLIPLOIDIA: FRUMENTO

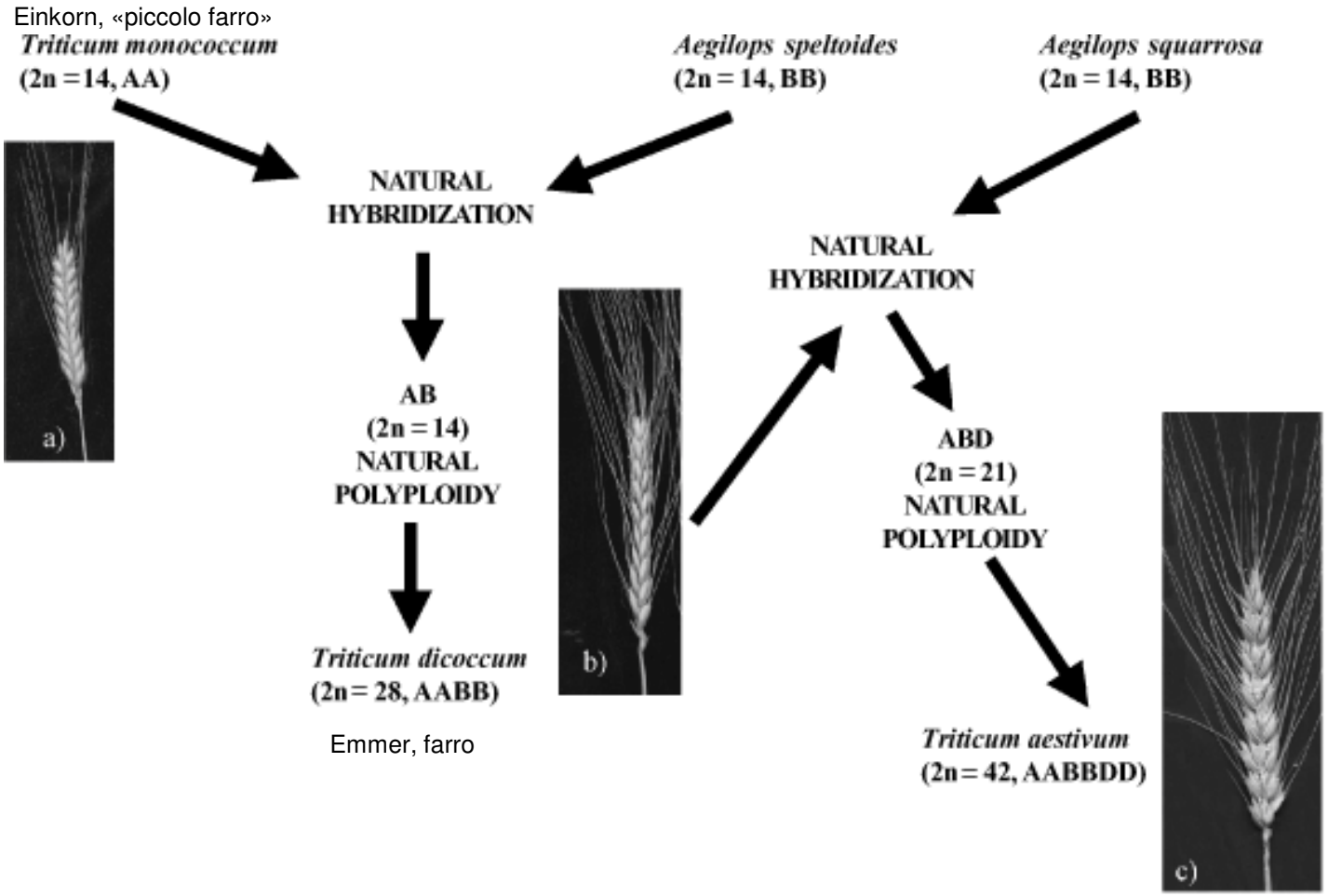
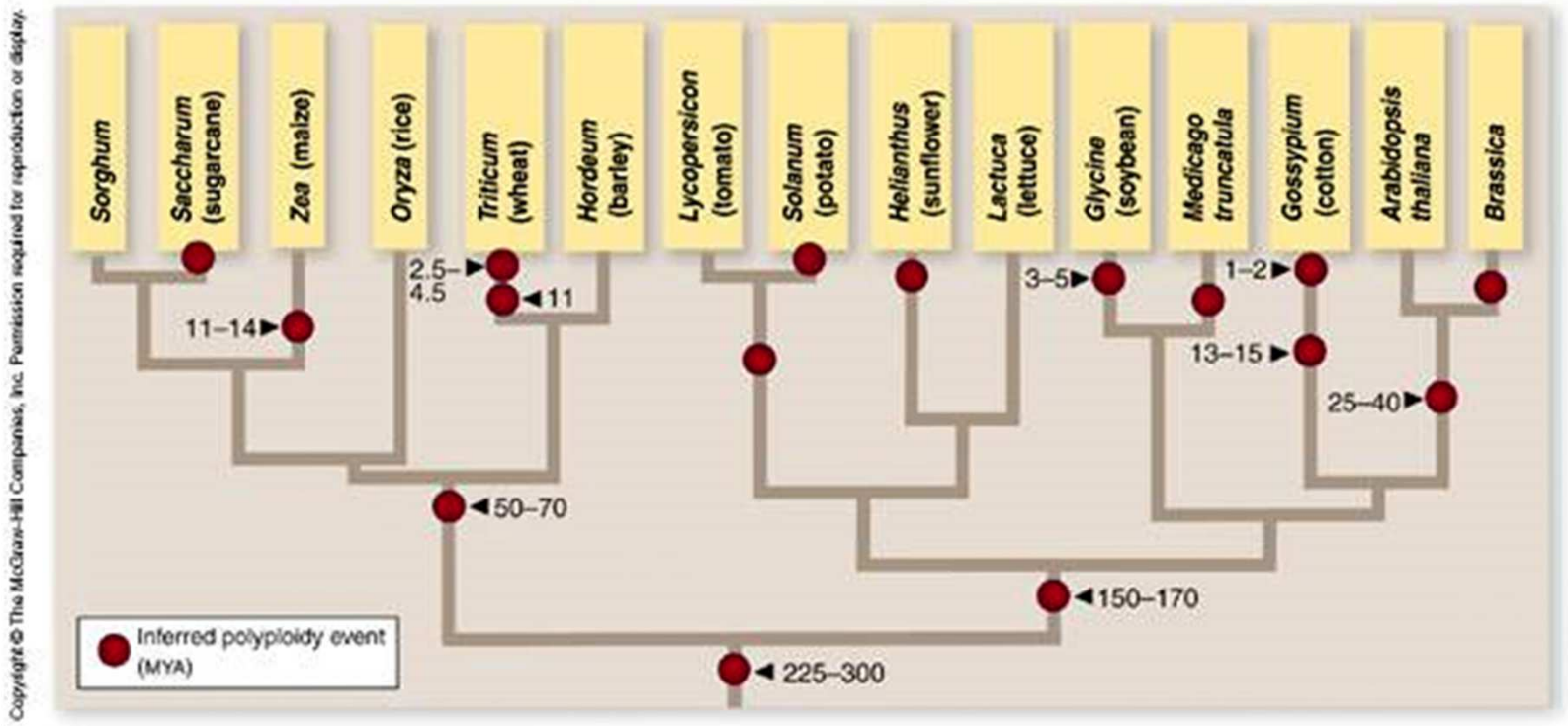


Figure 1 - Synoptic chart of cultivated wheats evolution: the diploid (2n = 14, AA) forms of *Triticum monococcum* (**a**) were naturally pollinated by weed species, possible *Aegilops speltoides* (2n = 14, BB?), in about 10,000 B.C. primitive farms. The subsequent genome duplication of hybrids by natural polyploidy gave rise to several wild and cultivated tetraploid species (2n = 28, AABB) like *Triticum dicoccum* (**b**) and *Triticum durum* (Figure 2a); again, the natural pollination of the tetraploid *T. dicoccum* (**b**) by another weed species, *Aegilops squarrosa* (2n = 14, DD) gave rise to the hexaploid (2n = 42, AABBDD) species (**c**).

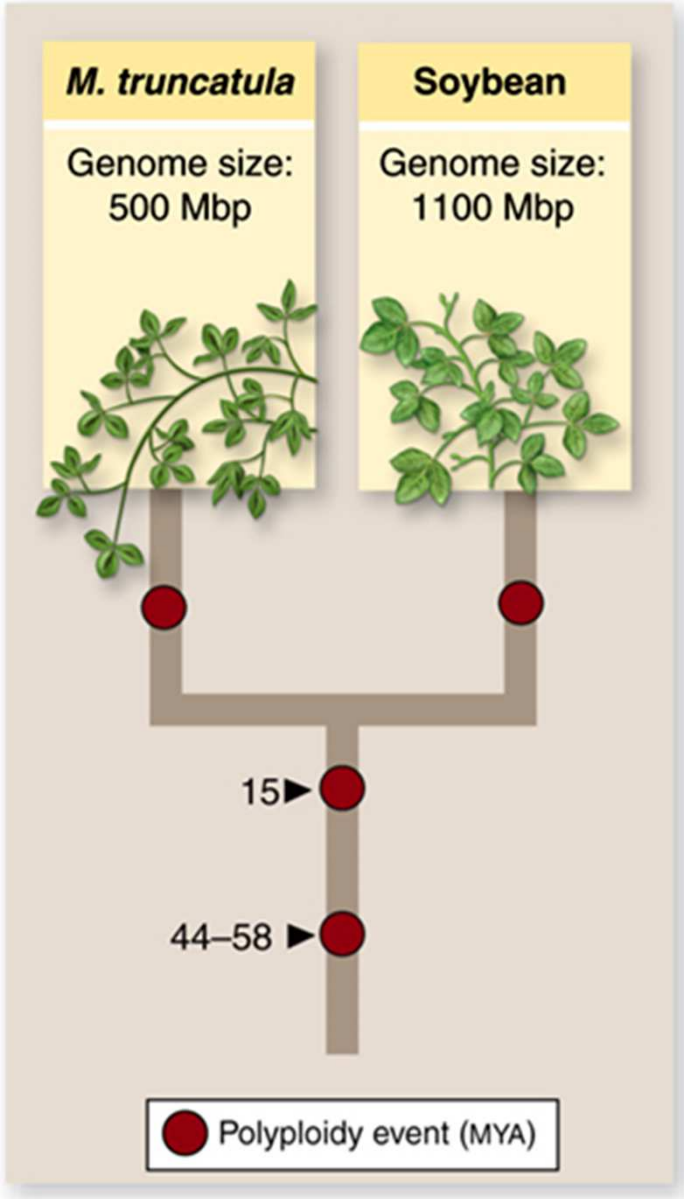
La poliploidia guida lo studio dell'evoluzione dei genomi

- **Paleopoliploidia:** confronto degli eventi di ploidizzazione
  - Divergenza di sequenze duplicate
  - Presenza o assenza di coppie di geni duplicati in seguito a ibridazione

Evoluzione dei genomi

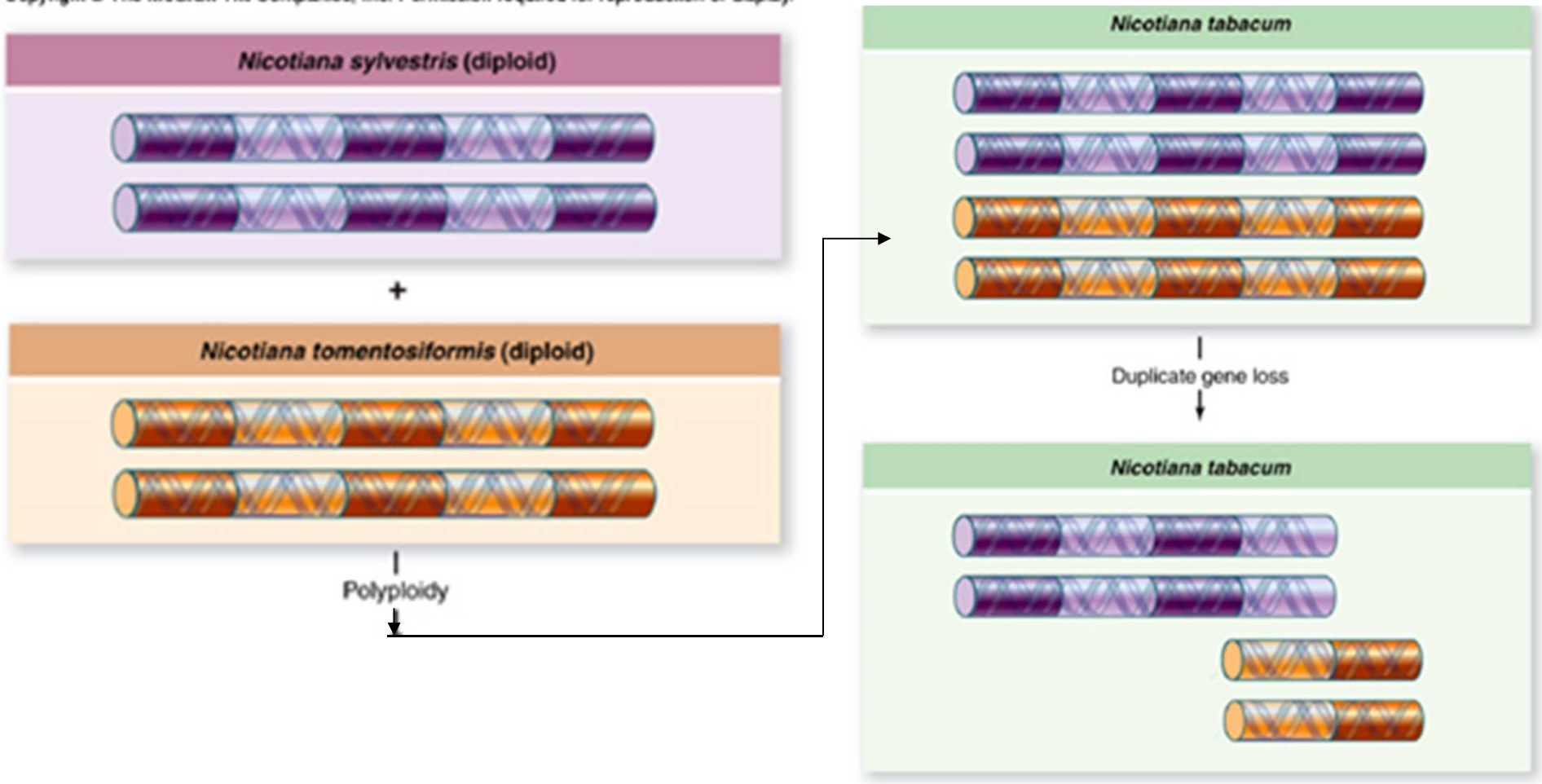


La poliploidia è diffusa nelle piante e ha avuto origini multiple durante l'evoluzione



Riduzione delle dimensioni del genoma

- Destino dei geni duplicati
  - Perdita di funzione per mutazione
  - Nuove funzioni
  - Suddivisione delle funzioni tra le due copie



Perdita di geni duplicati-> problema anche per identificare geni ortologhi in specie diverse

POLIPLOIDIA E SEQUENZIAMENTO DEI GENOMI

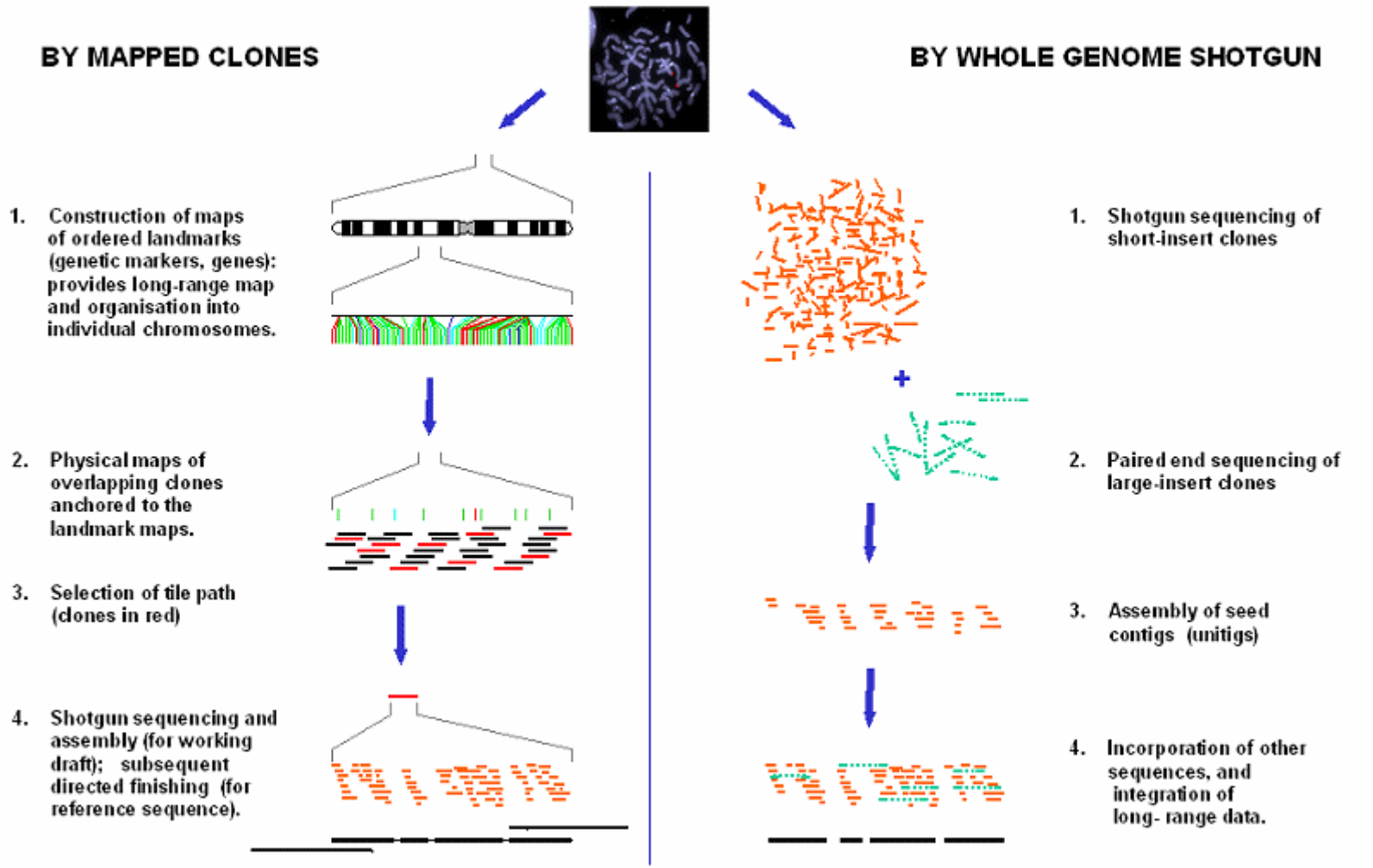
Molte specie autopoliploidi sono intolleranti all'INBREEDING, e hanno alti livelli di eterozigosità, importanti per la produttività

-> problema nell'assemblaggio dei contigs (più alleli diversi per ogni gene)

Negli allopoliploidi i cromosomi duplicati hanno subito sufficiente divergenza per non appaiarsi tra loro -> le sequenze delle coppie geniche sono distinguibili

N.B.: tutte le angiosperme sono PALEOPOLIPLOIDI, ma i geni "paleologhi" sono normalmente ben differenziati

Sequenziamento “whole-genome shotgun” o “clone-by-clone”?



“whole-genome shotgun”

Vantaggi

- Rapido
- Meno costoso
- Utile per sequenziare regioni refrattarie alla mappatura fisica (es. regioni ripetitive)

Svantaggi

- Assemblaggio complicato se ci sono molte regioni ripetitive
- In autopoliploidi, non distingue aplotipi diversi di geni identici

“clone-by-clone”

Vantaggi

- Delimita l’incertezza a intervalli piccoli (100Kb)
- Un allele alla volta -> no problema di eterozigosità

Svantaggi

- Costo dell’assemblaggio della library e dell’ordinamento dei contigs

EST = Expressed Sequence Tags

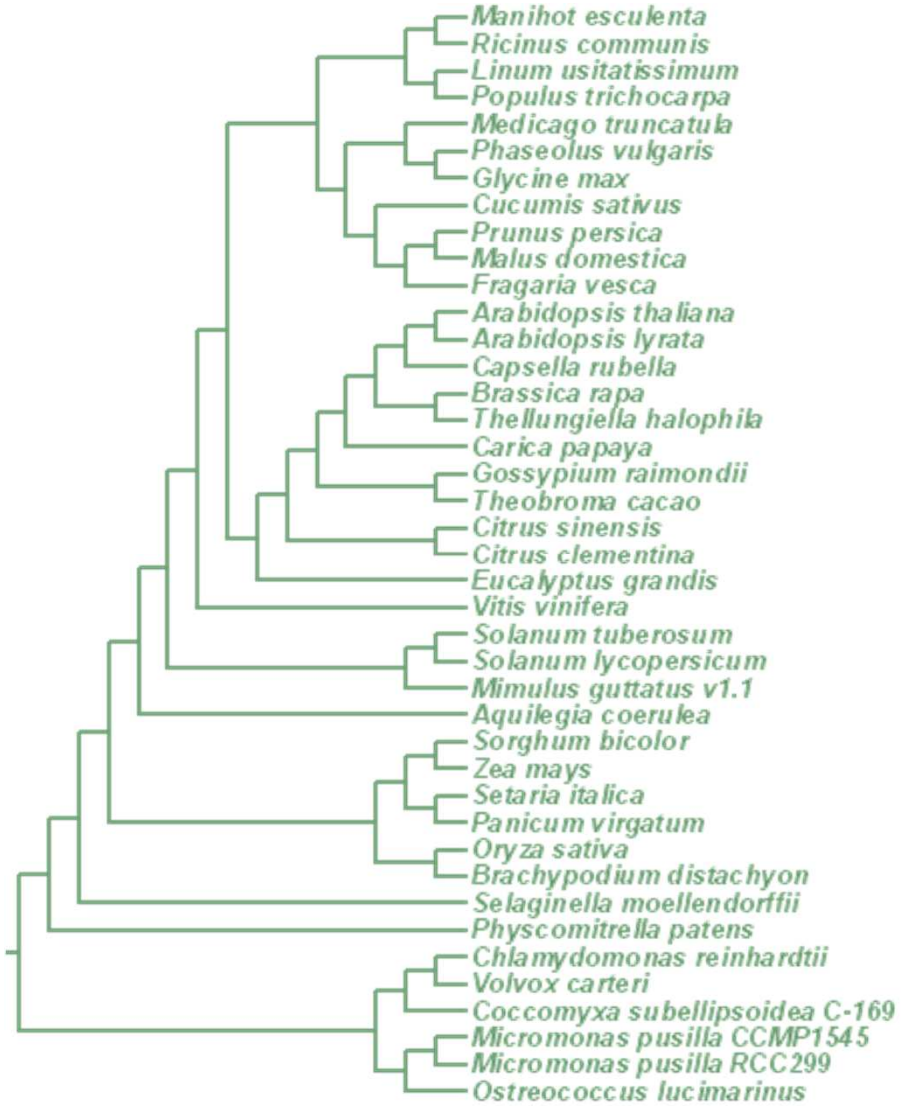
Creati sequenziando l'estremità 5' e/o 3' di mRNA isolati a caso e convertiti in cDNA (di solito 200–900 nt)

-> veloce e poco costoso
-> scoperta geni nuovi
 -> marcatori per mappatura
 -> base per futuri progetti di sequenziamento genomico
-> parziale copertura della porzione codificante del genoma

GENOMICA COMPARATIVA

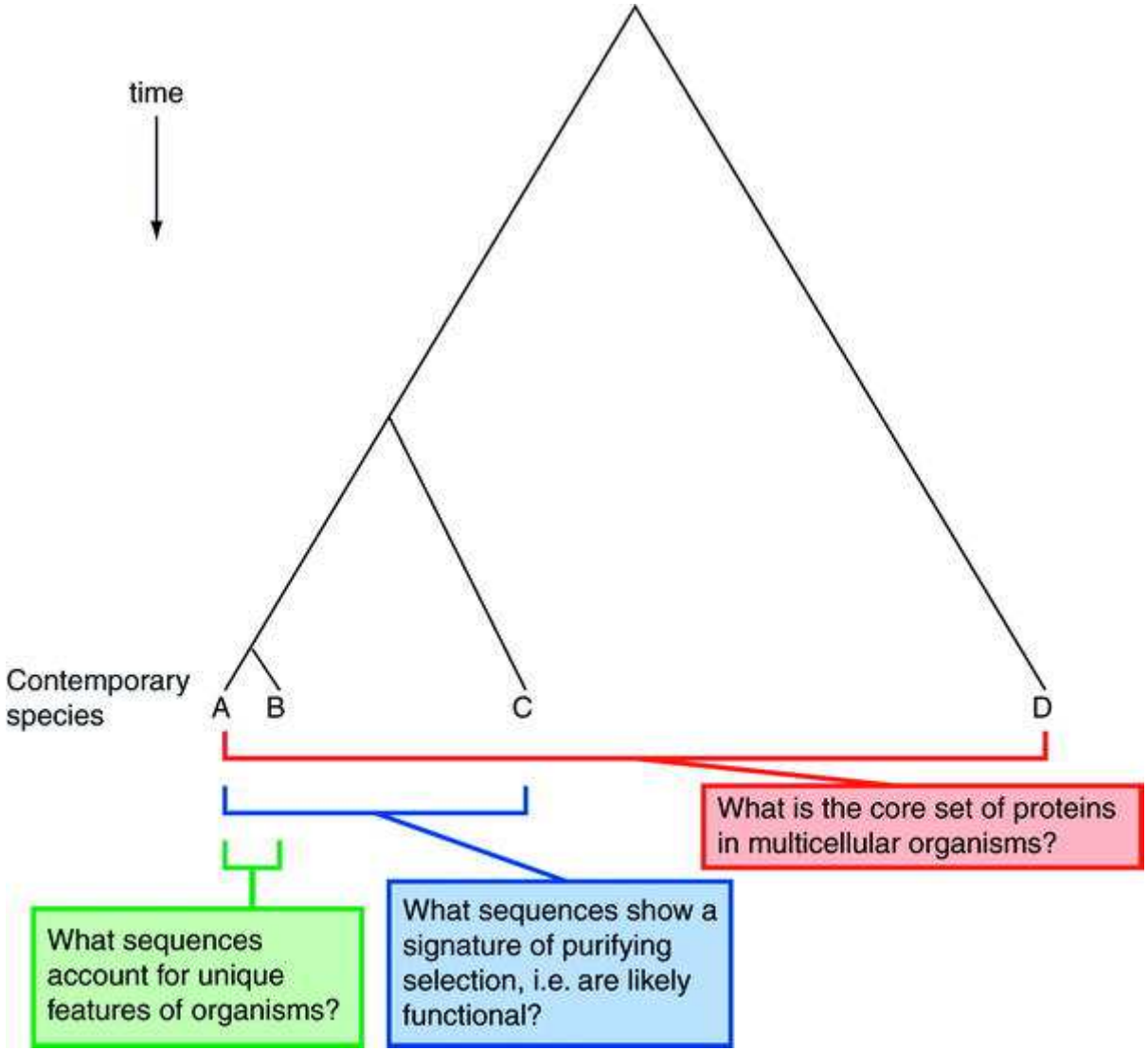
- Analisi e confronto di genomi di specie diverse
- Fornisce informazioni sull'evoluzione delle specie e sulla funzione di geni e sequenze non codificanti
- Es.: funzione di un gene dedotta dallo studio di geni ortologhi in specie modello

SEQUENCED AND ANNOTATED GREEN PLANT GENOMES



GENOMICA COMPARATIVA
Cosa si analizza?

- Similarità di sequenza
- Localizzazione cromosomica dei geni
- Lunghezza e numero esoni
- Quantità di DNA non codificante
- Conservazione di regioni cromosomiche


time
Contemporary species
A
B
C
D
What is the core set of proteins in multicellular organisms?
What sequences show a signature of purifying selection, i.e. are likely functional?
What sequences account for unique features of organisms?

Predizione della funzione di un gene a partire dalla sequenza di geni in altre specie

Gene con funzione ignota

Trasferimento di annotazione

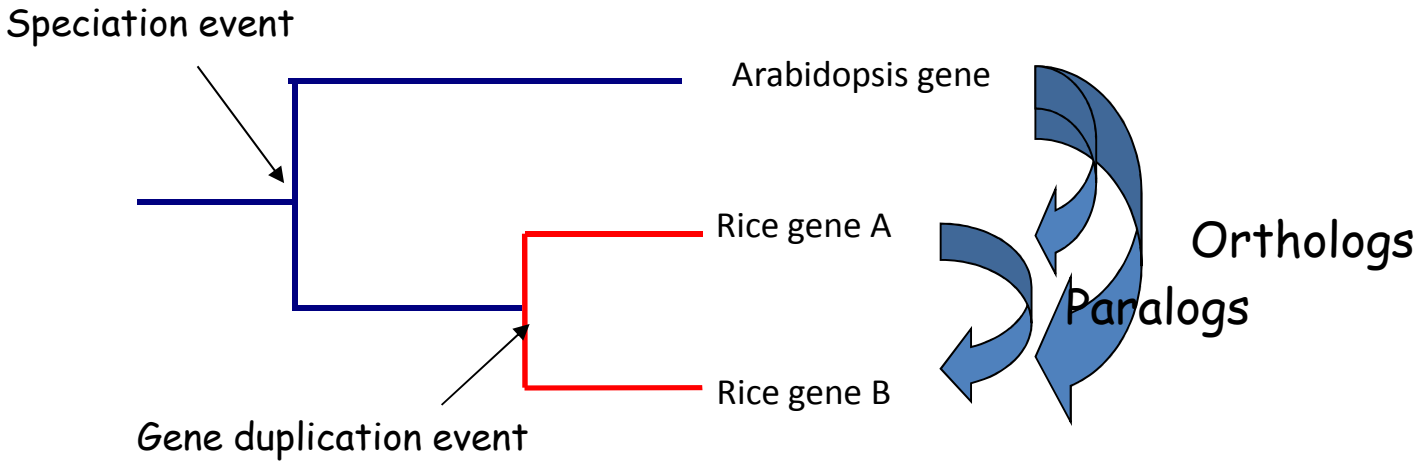
Specie modello

Gene con funzione X

Geni omologhi

Geni omologhi

- **Geni ortologhi** sono geni omologhi che discendono dall'ultimo ancestore comune attraverso speciazione
- Molto probabilmente codificano per proteine con funzione simile



- **Geni paraloghi** sono geni omologhi che si sono evoluti per duplicazione e possono codificare proteine con funzioni più divergenti
- **Geni inparaloghi**: geni ortologhi che hanno subito duplicazione

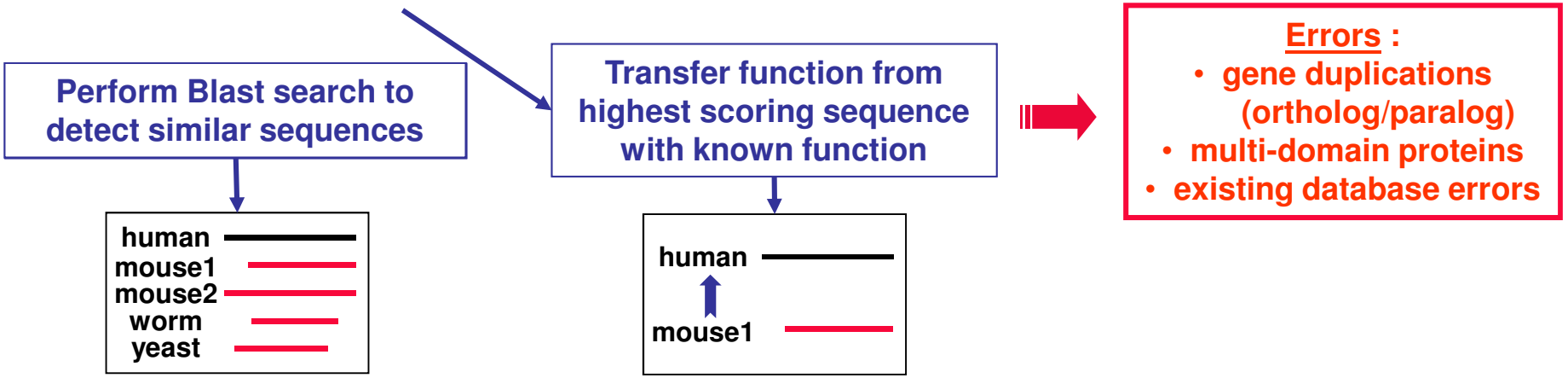
Come predirre l'omologia?

Similarità e omologia non sono la stessa cosa!

<u>Geni simili</u> si assomigliano sulla base di un'osservazione empirica

<u>Geni omologhi</u> sono geneticamente correlati (fatto storico: hanno antenato comune)

Metodo classico : annotazione funzionale basata sulla somiglianza (Blast)



Predizione dell'omologia sulla base della similarità

Es. BLAST

Vantaggi:

- **Facile**
- **Veloce**
- **Direttamente sul genoma completo**

Svantaggi:

- **Come stabilire la soglia di E-value per trasferire l'annotazione del gene da una specie all'altra?**

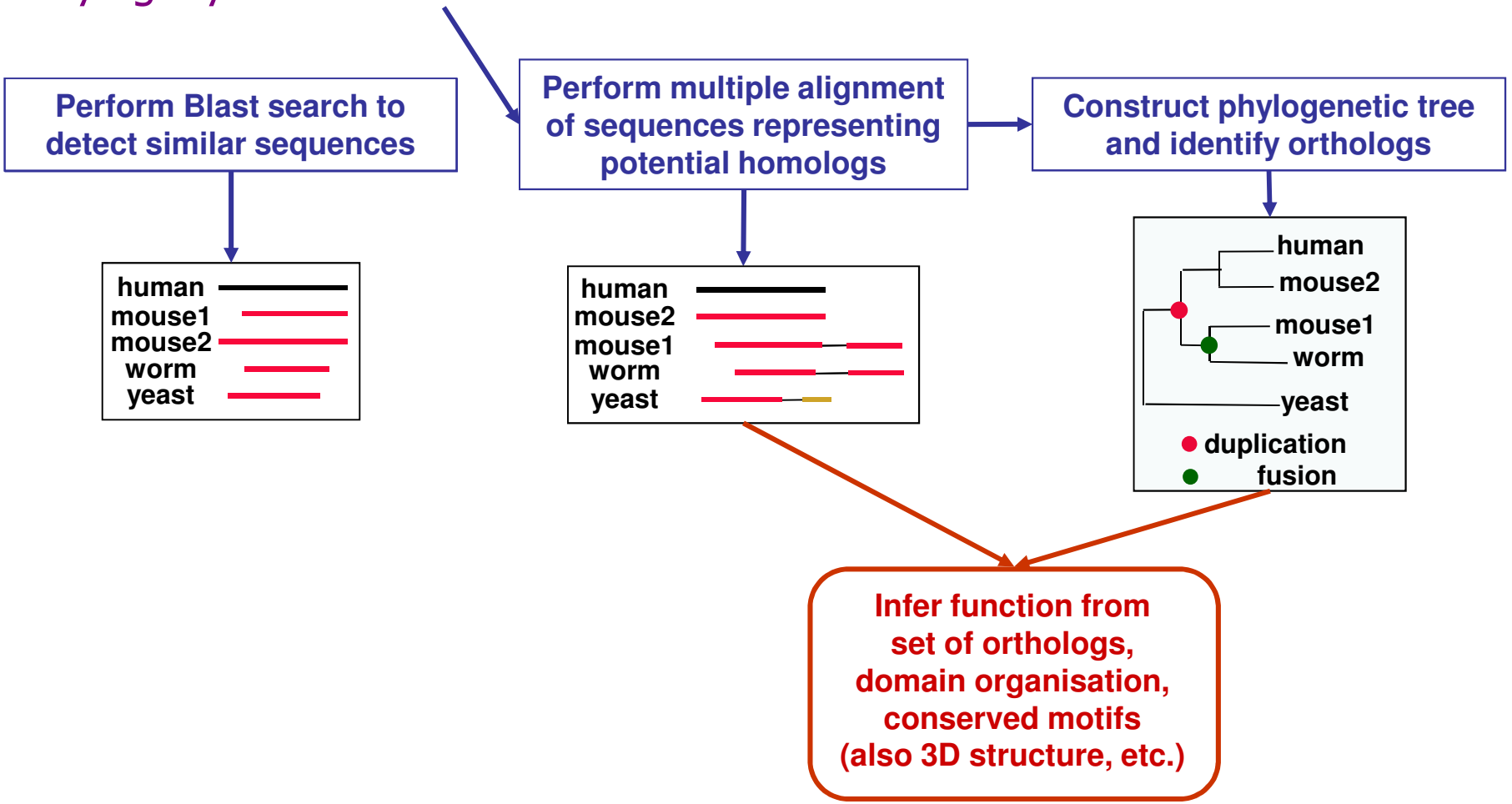
Due sequenze possono presentare similarità senza essere evolutivamente correlate!

- **Non identifica eventi di duplicazione genica**

Come trovare in una specie un gene ortologo ad un gene noto in un'altra specie?

⟶ **FILOGENOMICA**

Phylogeny-based inference



Predizione dell'omologia sulla base della filogenesi

Vantaggi:

- Efficiente per identificare duplicazioni (paraloghi e ortologhi)

Svantaggi:

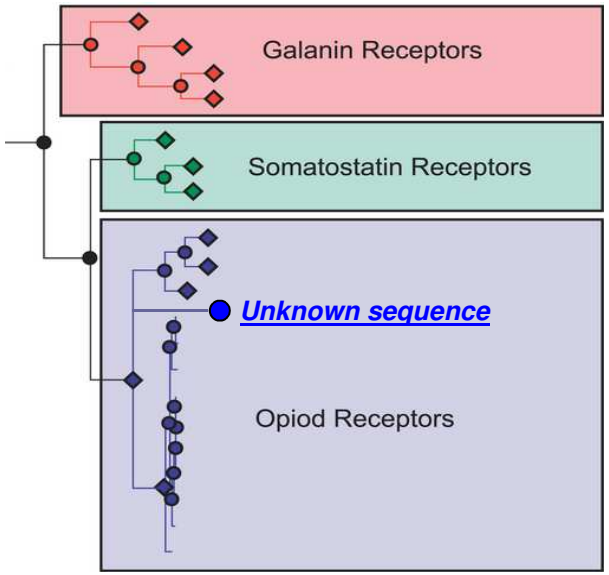
- Lento
- Richiede raggruppamento dei geni in famiglie

Metodi correnti

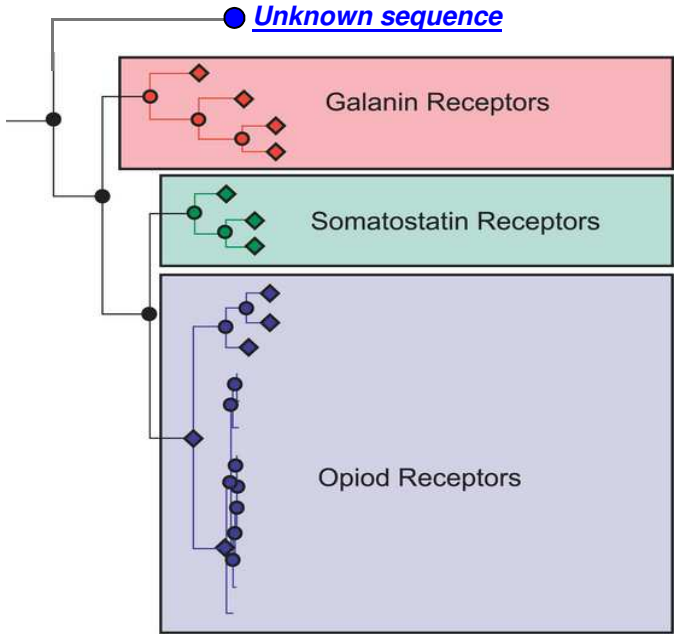
- RIO e Orthostrapper : solo per 1900 famiglie di geni vegetali (Pfam)
- GOST (usa GreenPhylDB family : 6420 famiglie geniche vegetali)

- Tree-based orthology: build a phylogenetic tree of a group of genes and compare gene tree to species tree to define speciation, duplication events
  - Resampled Inference of Orthologs (RIO) (Zmasek and Eddy, 2002)
  - Orthostrapper (Storm and Sonnhammer, 2002)
  - Levels Of Orthology From Trees (LOFT) (Van de Heijden et al, 2007)

- Example: G protein-coupled receptors



Prediction: Opiod receptor



More general prediction: GPCR of unknown specificity

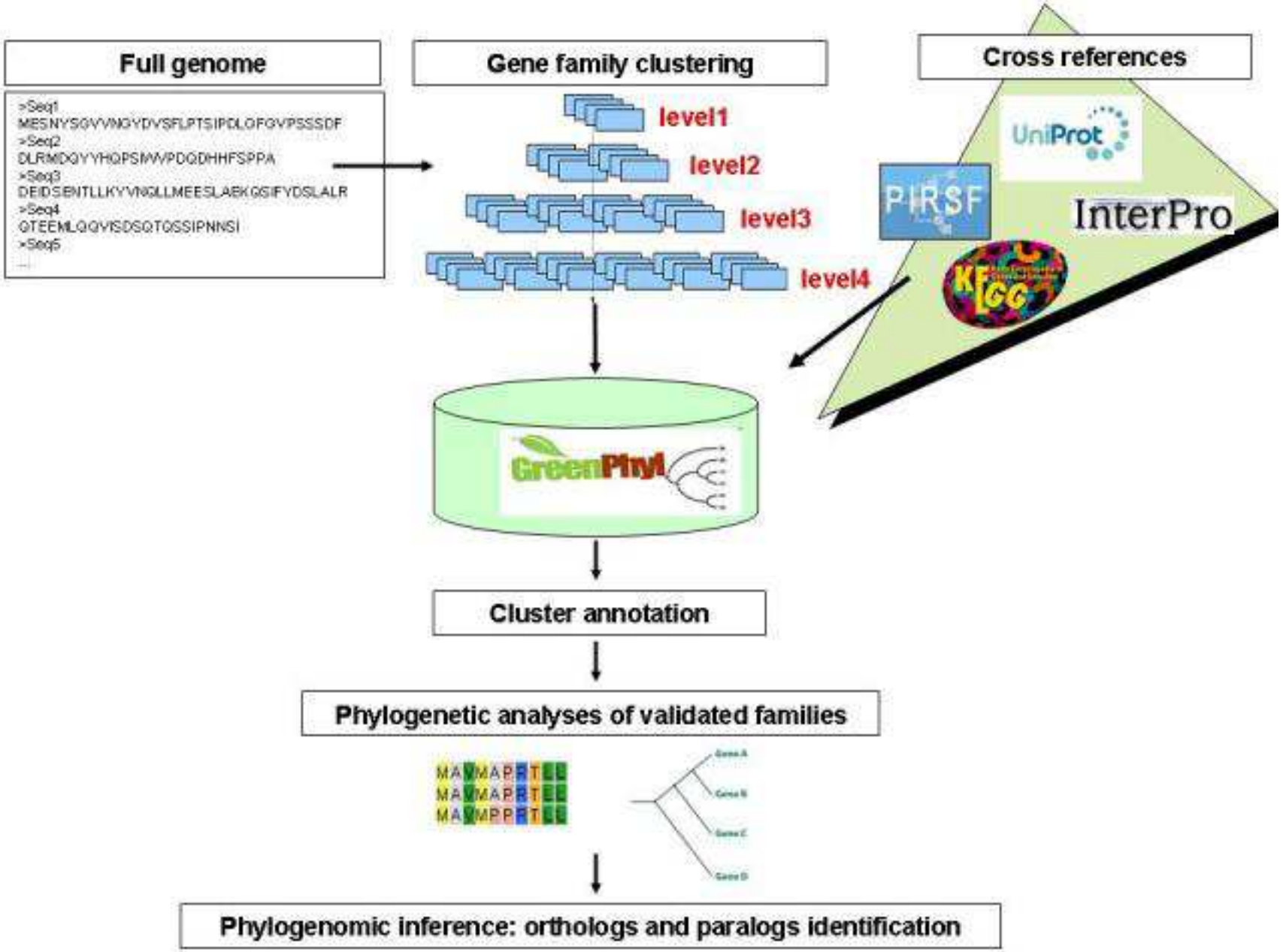
Due specie modello

Inizialmente *Oryza sativa* e *Arabidopsis thaliana:*

- **Genoma completo**
- **Alta qualità dell'annotazione (TAIR release 7, TIGR release 5)**
- **Evidenze funzionali disponibili**

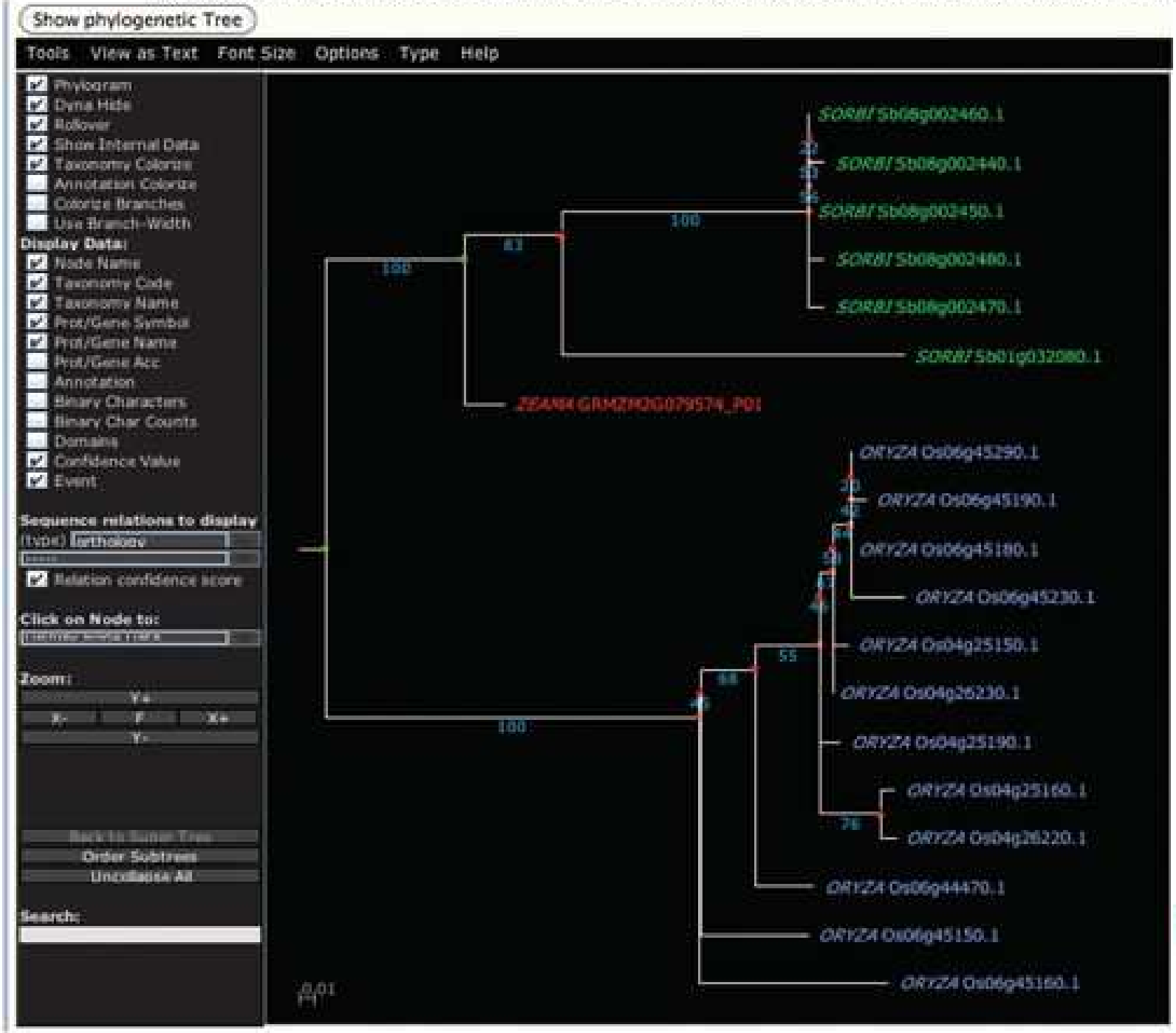
Nel tempo integra altre specie

http://www.greenphyl.org/cgi-bin/index.cgi


Full genome
>Seq1
MESNYSGVVNGYDVSFLPTSIPDLQFGVPSSSDF
>Seq2
DLRMDQYYHQPSMVVPDQDHHFSPPA
>Seq3
DEIDSENTLLKYVNGLLMEESLAEKGSIFYDSLALR
>Seq4
GTEEMLQGVISDSQTGSSIPNNSI
>Seq5
...
Gene family clustering
level1
level2
level3
level4
Cross references
UniProt
PIRSF
InterPro
KEGG
GreenPhyl
Cluster annotation
Phylogenetic analyses of validated families
MAVMAPRTEE
MAVMAPRTEE
MAVMPPRTEE
Phylogenomic inference: orthologs and paralogs identification

Amborella trichopoda	*Gossypium raimondii*	*Picea abies*
Arabidopsis thaliana	*Hordeum vulgare*	*Populus trichocarpa*
Brachypodium distachyon	*Lotus japonicus*	*Ricinus communis*
Cajanus cajan	*Malus domestica*	*Selaginella moellendorffii*
Carica papaya	*Manihot esculenta*	*Setaria italica*
Chlamydomonas reinhardtii	*Medicago truncatula*	*Solanum lycopersicum*
Cicer arietinum	*Musa acuminata*	*Solanum tuberosum*
Citrus sinensis	*Musa balbisiana*	*Sorghum bicolor*
Coffea canephora	*Oryza sativa*	*Theobroma cacao*
Cucumis sativus	*Ostreococcus tauri*	*Vitis vinifera*
Cyanidioschyzon merolae	*Phaseolus vulgaris*	*Zea mays*
Elaeis guineensis	*Phoenix dactylifera*	
Glycine max	*Physcomitrella patens*	


GreenPhyl
Family_name (e.g. GRAS)
Search
Home
Gene Family lists
Tools
Documentation
Statistics
About
Family ID 20923
Family name Pollen Allergen/Expansin Superfamily
Synonym(s)
Cross-reference(s)
Curation status
Phylogenetic analyzes Not available
(a)
Family structure
Family composition
Protein domains
Protein list
Phylogenomic analysis
Chromosome position
Clustering level
Family id (number of sequences)
1 20923 (812)
2 24999 (689) 25670 (26) 26339 (66)
3 30583 (515) 31138 (117) 31147 (26) 31717 (57) 31804 (66)
4 36255 (410) 36521 (194) 36760 (26) 37177 (65) 37759 (61)
(b)
Family structure
Family composition
Protein domains
Protein list
Phylogenomic analysis
Chromosome position
InterPro family
IPR Annotation Type % Occurence Specificity
IPR007118 Expansin/Lol pI Family 80 (653) Y
Other InterPro signatures
IPR Annotation Type % Occurence Specificity
IPR014734 Pollen allergen, N-terminal Domain 57 (466) N
IPR009009 Barwin-related endoglucanase Domain 58 (472) N
IPR005132 Rare lipoprotein A Domain 89 (722) N
IPR007117 Pollen allergen/expansin, C-terminal Domain 82 (667) N
IPR007112 Expansin 45, endoglucanase-like Domain 88 (714) N
Domain architecture
Identified using MEME suite
see sequence motifs
Representative InterPro domains - Consensus schema (alpha version)


Show phylogenetic Tree
Tools
View as Text
Font Size
Options
Type
Help
Phylogram
Dyna Hide
Rollover
Show Internal Data
Taxonomy Colorize
Annotation Colorize
Colorize Branches
Use Branch-Width
Display Data:
Node Name
Taxonomy Code
Taxonomy Name
Prot/Gene Symbol
Prot/Gene Name
Prot/Gene Acc
Annotation
Binary Characters
Binary Char Counts
Domains
Confidence Value
Event
Sequence relations to display
(type) orthology
Relation confidence score
Click on Node to:
Zoom:
Y+
X-
F
X+
Y-
Order Subtrees
Uncollapse All
Search:
SORBI Sb06g002460.1
SORBI Sb06g002440.1
SORBI Sb06g002450.1
SORBI Sb06g002480.1
SORBI Sb06g002470.1
SORBI Sb01g032080.1
ZEAMA GRMZM0G079574_P01
ORYZA Os06g45290.1
ORYZA Os06g45190.1
ORYZA Os06g45180.1
ORYZA Os06g45230.1
ORYZA Os04g25150.1
ORYZA Os04g26230.1
ORYZA Os04g25190.1
ORYZA Os04g25160.1
ORYZA Os04g26220.1
ORYZA Os06g44470.1
ORYZA Os06g45150.1
ORYZA Os06g45160.1
100
83
100
100
55
68
76
0.01

(e)

InParanoid

- http://inparanoid.sbc.su.se

- Database per identificare geni ortologhi e inparaloghi tra specie diverse di eucarioti (animali, piante, funghi, protisti)