Sequenziamento

Prima generazione - metodo sviluppato da Sanger nel 1980

Seconda generazione (next generation sequencing, NGS) – Pirosequenziamento

– Illumina

– SOLiD (sequencing by oligo ligation and detection)

Terza generazione (next next generation sequencing, NNGS)- Helicos

- PacBio

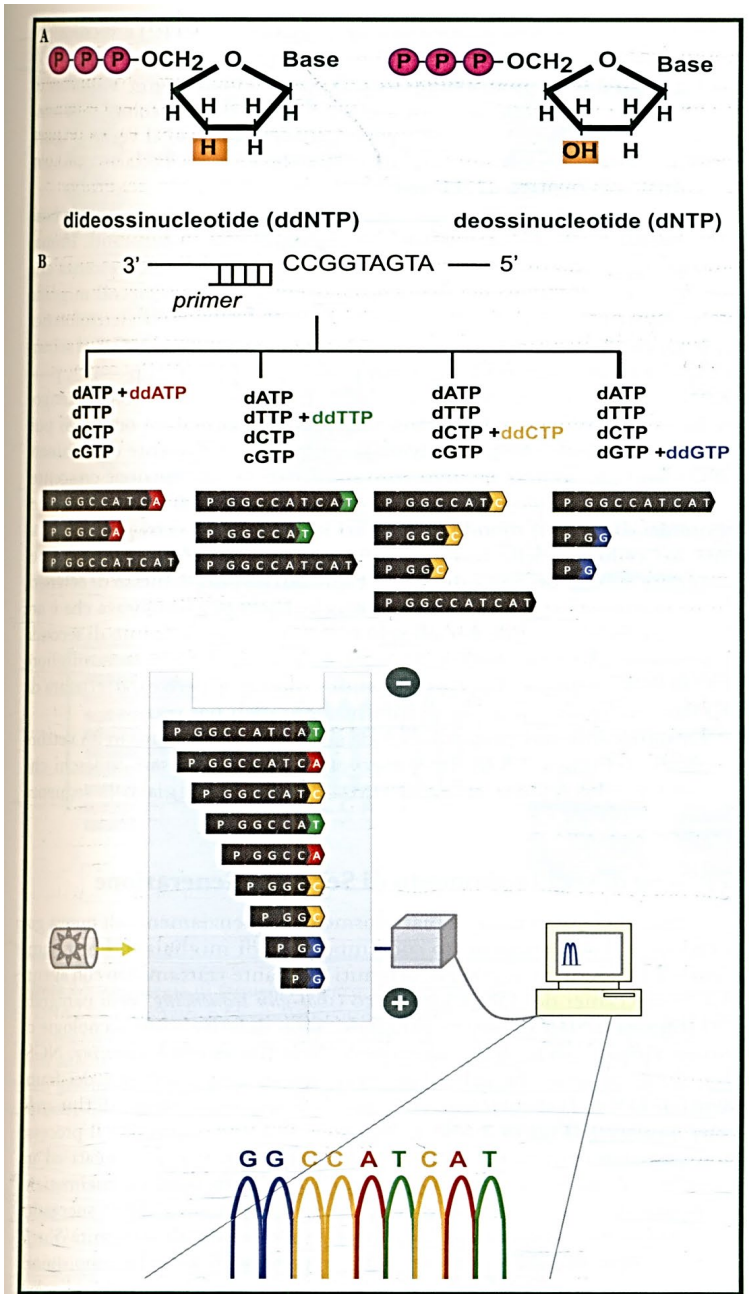
- Nanopore

- Ion Torrent

Prima generazione - metodo Sanger o a terminazione di catena

1) Primer
2) DNA polimerasi
3) dNTP
4) ddNTP nucleotidi modificati privi del gruppo ossidrile al 3’ interrompono l’allungamento della catena

Ottenimento di frammenti amplificati di diversa lunghezza

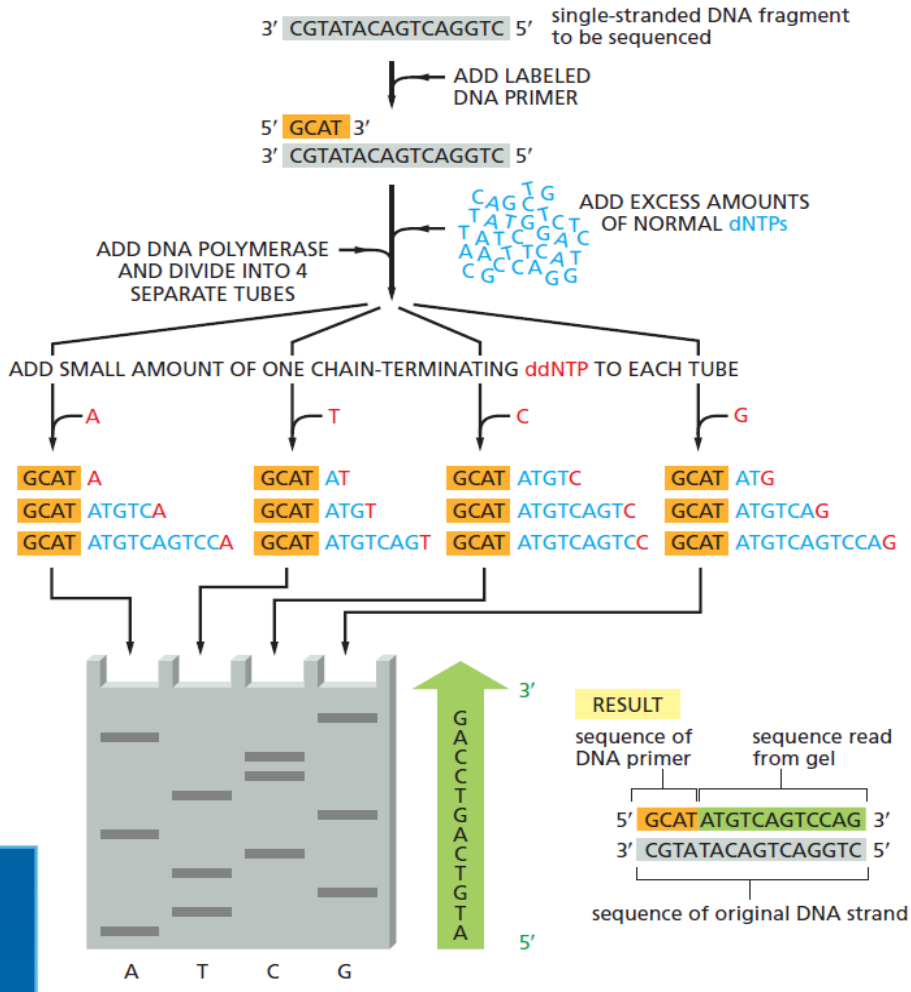


L’elettroforesi li ordina in base alla lunghezza e il detector riconosce i 4 ddNTP poiché marcati con 4 diversi fluorocromi

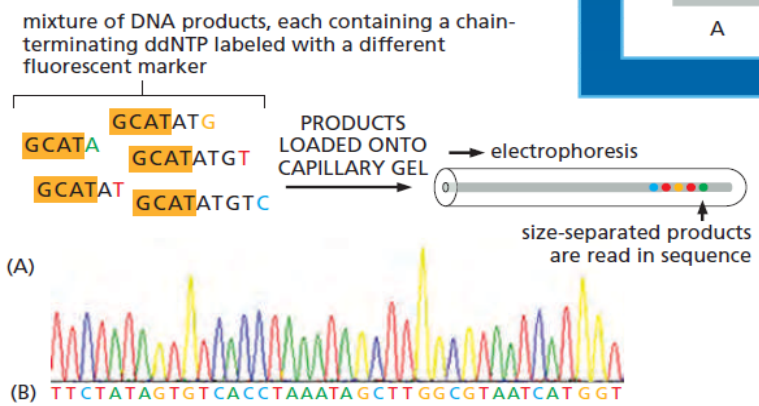
-Migliore qualità delle sequenze
-Si possono ottenere sequenze che superano le 1000bp
-Bassa velocità
-Bassa quantità di informazioni

MANUAL DIDEOXY SEQUENCING

To determine the complete sequence of a single-stranded fragment of DNA *(gray)*, the DNA is first hybridized with a short DNA primer *(orange)* that is labeled with a fluorescent dye or radioisotope. DNA polymerase and an excess of all four normal deoxyribonucleoside triphosphates (*blue* A, C, G, or T) are added to the primed DNA, which is then divided into four reaction tubes. Each of these tubes receives a small amount of a single chain-terminating dideoxyribonucleoside triphosphate (*red* A, C, G, or T). Because these will be incorporated only occasionally, each reaction produces a set of DNA copies that terminate at different points in the sequence. The products of these four reactions are separated by electrophoresis in four parallel lanes of a polyacrylamide gel (labeled here A, T, C, and G). In each lane, the bands represent fragments that have terminated at a given nucleotide but at different positions in the DNA. By reading off the bands in order, starting at the bottom of the gel and reading across all lanes, the DNA sequence of the newly synthesized strand can be determined (see Figure 8–25C). The sequence, which is given in the green arrow to the right of the gel, is complementary to the sequence of the original *gray* single-stranded DNA.

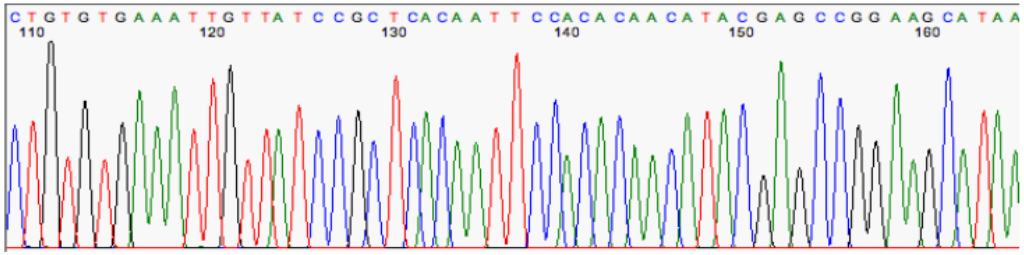


AUTOMATED DIDEOXY SEQUENCING

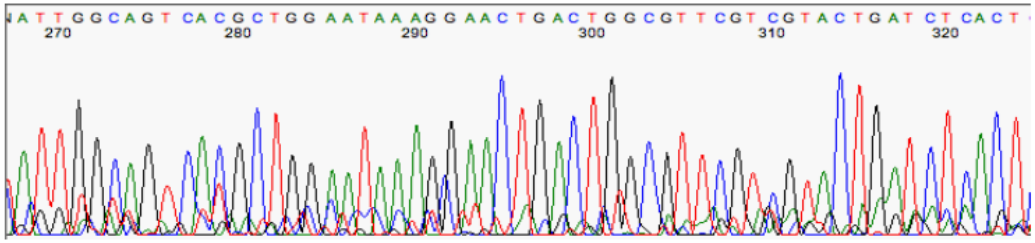


Fully automated machines can run dideoxy sequencing reactions. (A) The automated method uses an excess amount of normal dNTPs plus a mixture of four different chain-terminating ddNTPs, each of which is labeled with a fluorescent tag of a different color. The reaction products are loaded onto a long, thin capillary gel and separated by electrophoresis. A camera (not shown) reads the color of each band as it moves through the gel and feeds the data to a computer that assembles the sequence. (B) A tiny part of the data from such an automated sequencing run. Each colored peak represents a nucleotide in the DNA sequence.

Il risultato di un sequenziamento è visualizzato sotto forma di cromatogramma, che visualizza le emissioni in fluorescenza che identificano ciascuna delle 4 basi del DNA. Ciò risulta in una serie di picchi di fluorescenza (a frequenze diverse), che possono essere ben definiti, come nel cromatogramma che segue, privo di rumore di fondo:



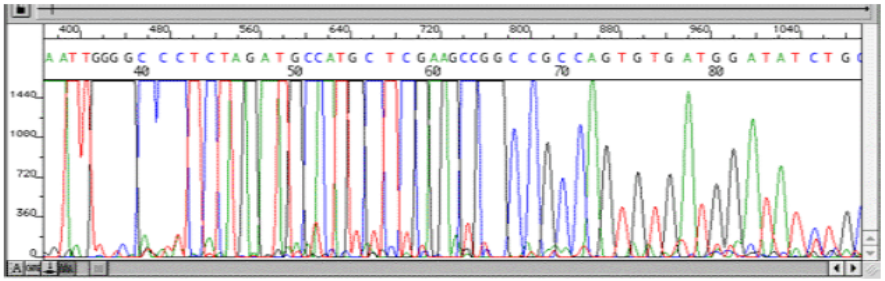
oppure possono presentare rumore di fondo:



Si può notare che alcuni picchi (ad es. quelli nelle posizioni 271, 273 e 279) sono sovrapposti; inoltre, c'è un picco a cavallo delle posizioni 291 e 292 ed in posizione 310 c'è una forte sovrapposizione di picchi.

Alle estremità 5' e 3' del cromatogramma si trovano quasi sicuramente sequenze di bassa qualità:

5' terminale

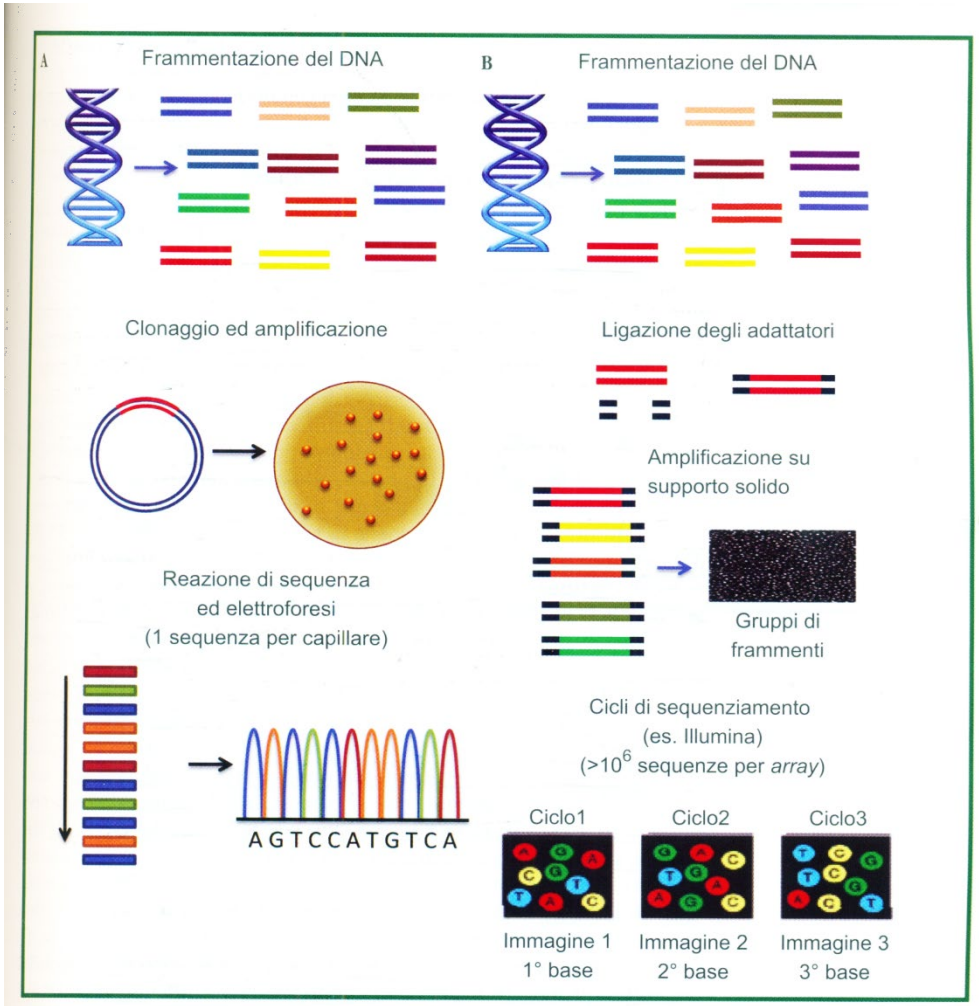


All'inizio della reazione di sequenziamento, i frammenti di DNA finiscono con il terminatore fluorescente, come ci si aspetta, ma sono molto corti; questo favorisce una concentrazione eccessiva dei frammenti e un effetto di "overload" del segnale.

Prima e seconda generazione

Metodo Sanger-
il frammento da sequenziare è clonato in un vettore plasmidico
-possibilità di appaiare gli inneschi in regioni note del plasmide e fiancheggianti la sequenza clonata

-Migliore qualità delle sequenze
-Si possono ottenere sequenze che superano le 1000bp
-Bassa velocità
-Bassa quantità di informazioni



Sequenziamento contemporaneo di migliaia di brevi frammenti (200bp circa)
- Ligazione di adattatori
- Amplificazione su supporto solido

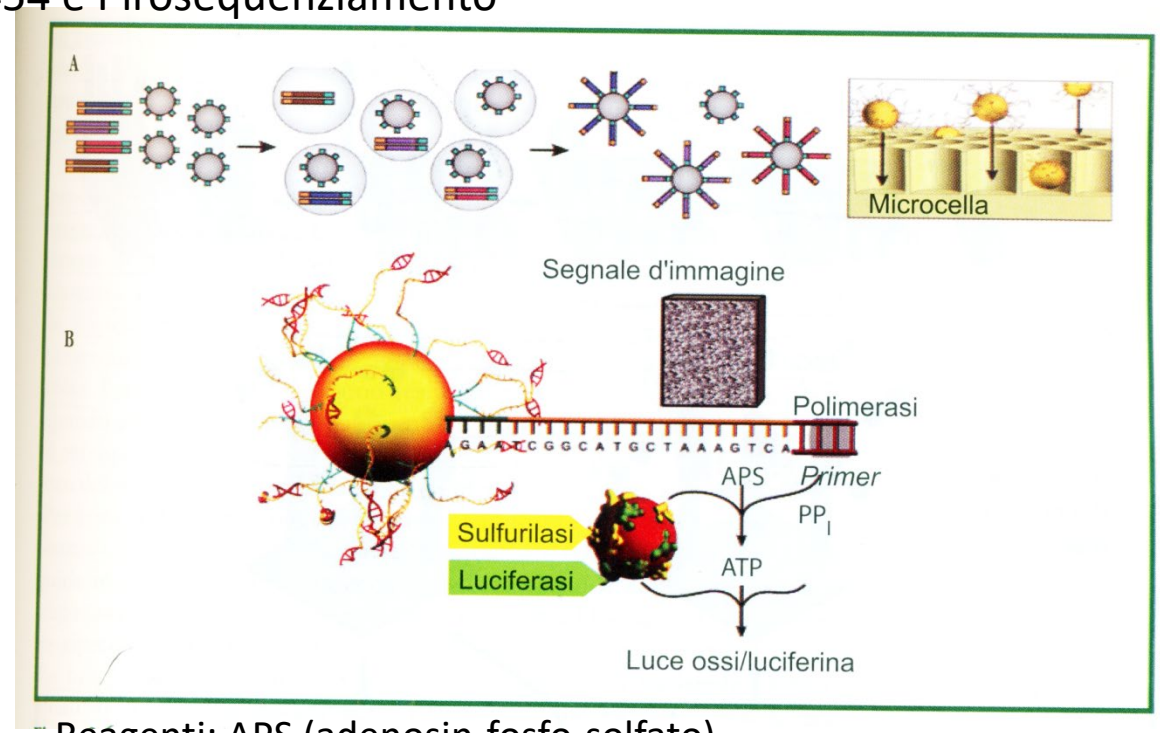
-Ottenimento contemporaneo di un numero elevato di sequenze su piccole superfici
-Basso costo
-Limitata lunghezza delle sequenze
-Accuratezza inferiore rispetto al metodo Sanger
-Necessità d strumenti informatici evoluti

Seconda generazione (next generation sequencing, NGS)

Tecnologia 454 e Pirosequenziamento

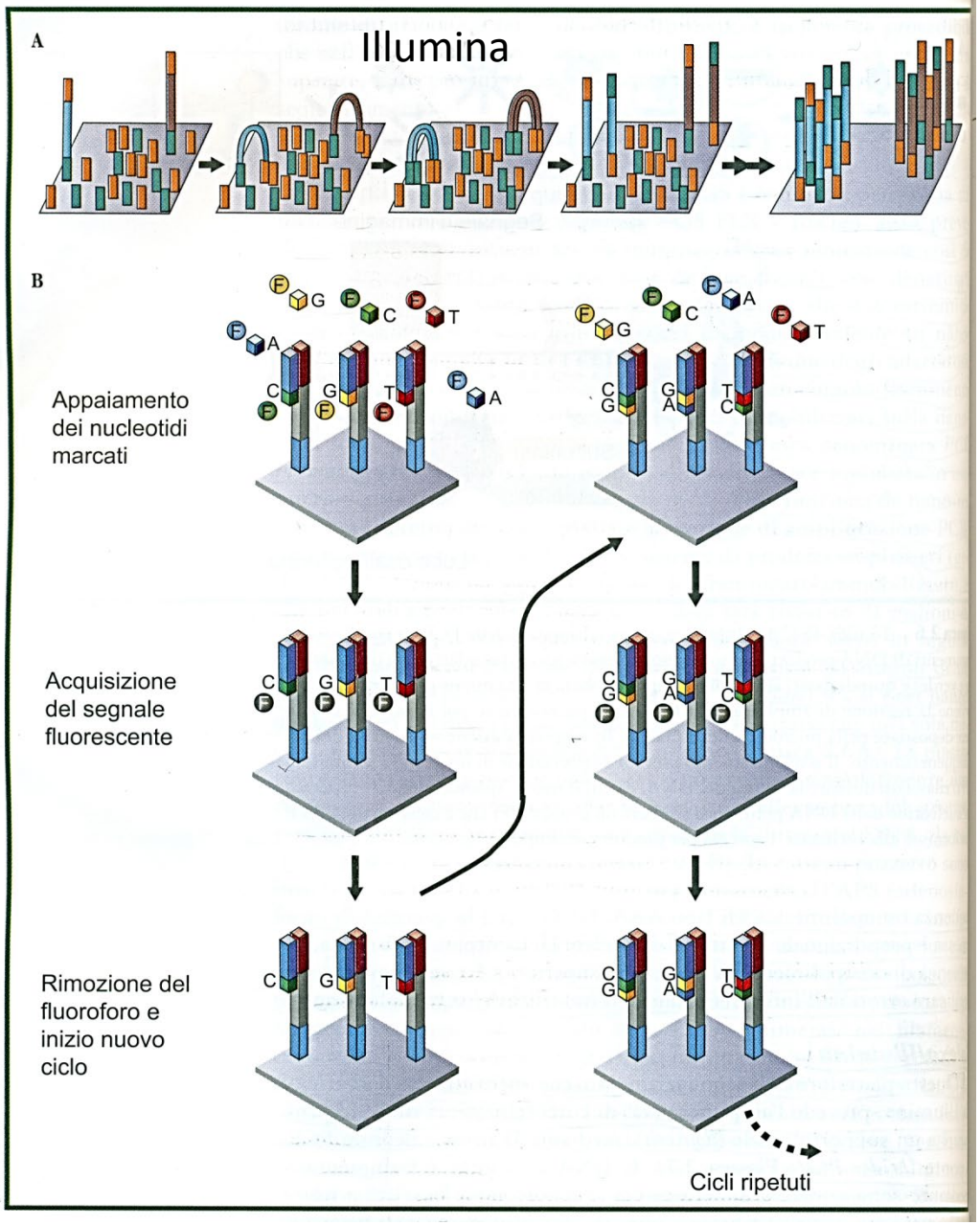
-Frammentazione del DNA per sonicazione- frammenti da 300-800 bp
-Adattatori legati alle estremità
-Denaturazione
-Immobilizzazione su nanosfere nelle quali è legata la sequenza di uno degli adattatori
-Un frammento per nanosfera
-Amplificazione attraverso PCR in emulsione
-Micro gocce che agiscono da reattore
-Una sfera in un poro 44 µm

Sequenziamento basato sulla rilevazione del gruppo pirofosfato rilasciato durante la sintesi di DNA



Reagenti: APS (adenosin-fosfo-solfato)
DNA polimerasi
ATP- sulfurilasi
Luciferina
Luciferasi
dNTP

Quando viene incorporato un nucleotide dalla DNA polimerasi si libera il gruppo Ppi che l'enzima sulfurilasi converte in ATP che fornisce energia alla luciferasi per ossidare il substrato che a questo punto da fluorescenza che viene rilevata

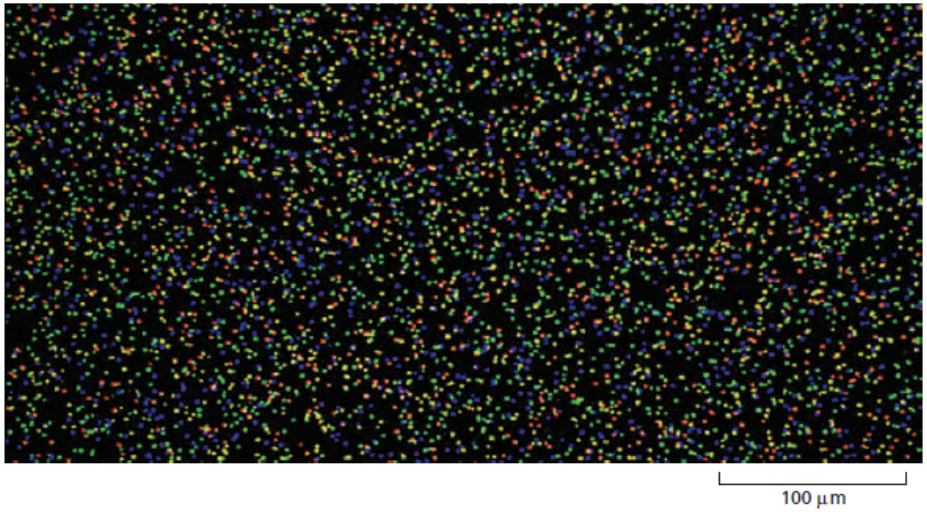


Frammenti di DNA ancorati su supporto solido amplificati mediante PCR a ponte

-Frammenti legati a due adattatori alle estremità
-Ogni adattatore è complementare ai nucleotidi fissati al supporto solido
-Denaturazione
-Ibridazione con i primers del supporto solido
-Dal secondo ciclo i filamenti si incurvano e si appaiano con i primers fissati sul supporto solido
-Si formano quindi dei cluster (1000 copie circa) isolati spazialmente che viene sottoposto a sequenziamento denominato 'base per base'

4 nucleotidi marcati con 4 diversi fluorofori modificati al 3' per evitare l'inserimento di più di un nucleotide per volta

Sequenze lunghe 150bp errori dovuti alla possibile sostituzione dei nucleotidi ad opera della polimerasi

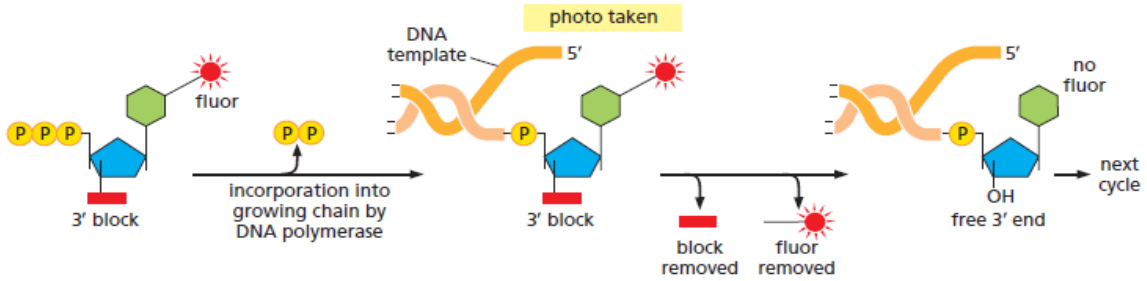


A slide showing individual clusters of PCR-generated DNA molecules. Each cluster carries about 1000 identical DNA molecules; the four colors are produced by incorporation of C, G, A, or T, each of which has a different color fluorophore. The image has been taken just after a fluorescent nucleotide has been incorporated into each growing DNA chain. (From Illumina Sequencing Overview, 2013.)

ILLUMINA® SEQUENCING

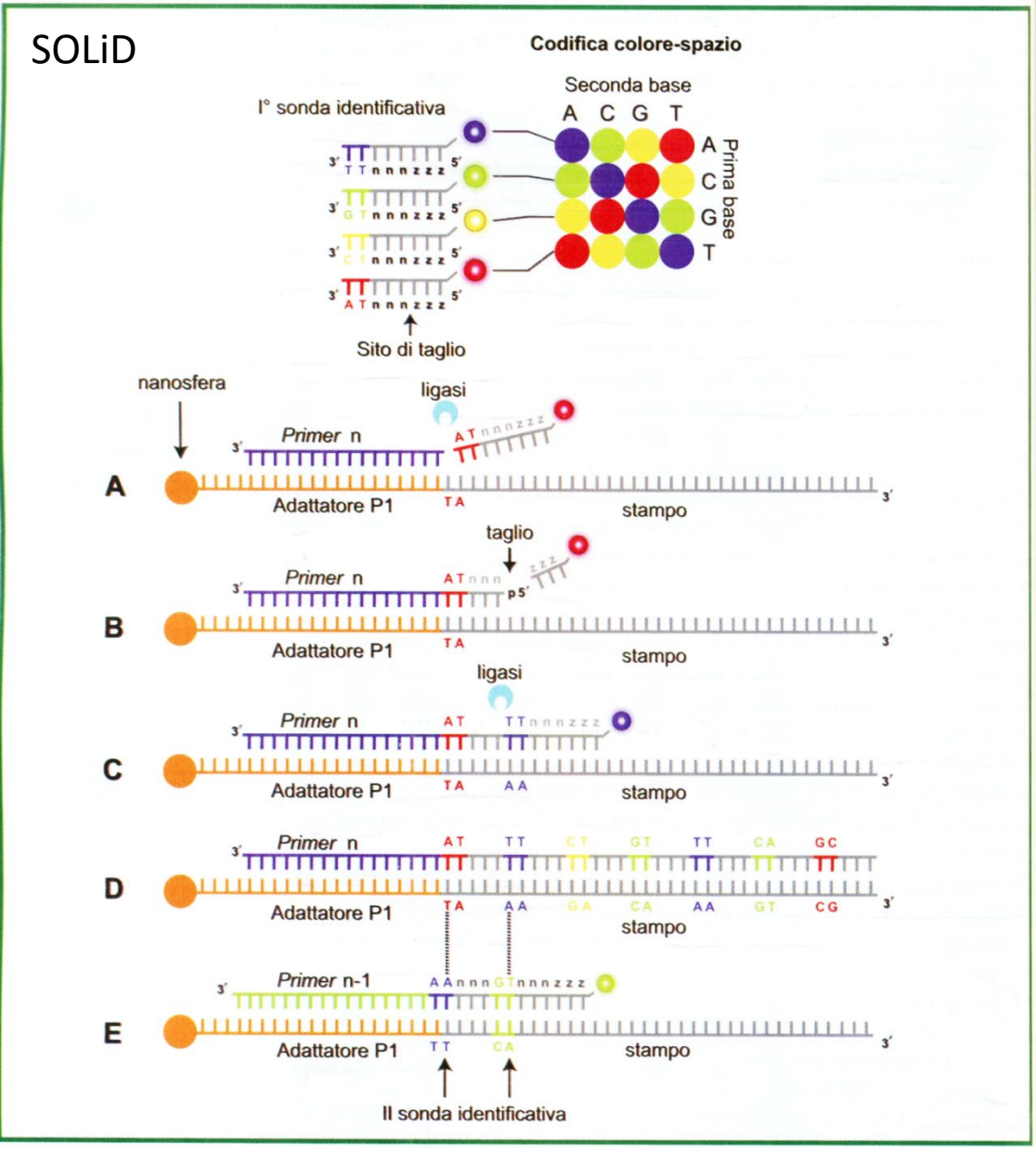
Several second-generation sequencing methods are now in wide use, and we will discuss two of the most common. Both rely on the construction of libraries of DNA fragments that represent—*in toto*—the DNA of the genome. Instead of using bacterial cells to generate these libraries, as we saw in Figure 8–30), they are made using PCR amplification of billions of DNA fragments, each attached to a solid support. The amplification is carried out so that the PCR-generated copies, instead of floating away in solution, remain bound in proximity to the original DNA fragment. This process generates clusters of DNA fragments, where each cluster contains about 1000 identical copies of a small bit of the genome. These clusters—a billion of which can fit in a single slide or plate—are then sequenced at the same time; that is, in parallel.

One method, known as *Illumina sequencing*, is based on the dideoxy method described above, but it incorporates several innovations. Here, each nucleotide is attached to a removable fluorescent molecule (a different color for each of the four bases) as well as a special chain-terminating chemical adduct: instead of a 3′-OH group, as in conventional dideoxy sequencing, the nucleotides carry a chemical group that blocks elongation by DNA polymerase but which can be removed chemically. Sequencing is then carried out as follows: the four fluorescently labeled nucleotides along with DNA polymerase are added to billions of DNA clusters immobilized on a slide. Only the appropriate nucleotide (that is complementary to the next nucleotide in the template) is covalently incorporated at each cluster; the unincorporated nucleotides are washed away, and a high-resolution digital camera takes an image that registers which of the four nucleotides was added to the chain at each cluster. The fluorescent label and the 3′-OH blocking group are then removed enzymatically, washed away, and the process is repeated many times. In this way, billions of sequencing reactions are carried out simultaneously. By keeping track of the color changes occuring at each cluster, the DNA sequence represented by each spot can be read. Although each individual sequence read is relatively short (approximately 200 nucleotides), the billions that are carried out simultaneously can produce several human genomes worth of sequence in about a day.



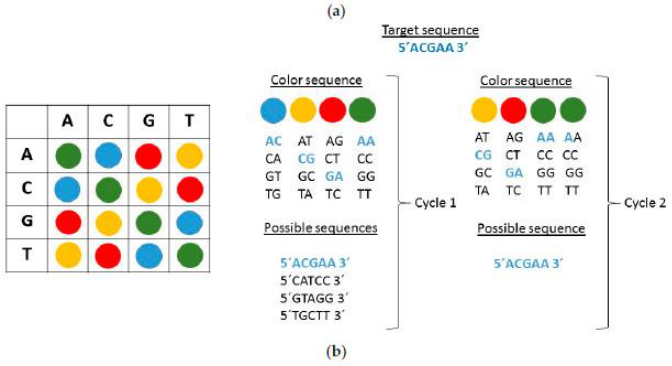
Principle behind Illumina sequencing. This reaction is carried out stepwise, on billions of DNA clusters at once. The method relies on a color digital camera that rapidly scans all the DNA clusters after each round of modified nucleotide incorporation. The DNA sequence of each cluster is then determined by the sequence of color changes it undergoes as the elongation reaction proceeds stepwise. Each round of modified nucleotide incorporation, image acquisition, and removal of the 3′ block and the fluorescent group takes less than an hour. Each cluster on the slide contains many copies of different, random bits of a genome; in preparing the clusters, a DNA sequence (specified by the experimenter) is joined to each copy in every cluster, and a primer complementary to this sequence is used to begin the elongation reaction by DNA polymerase.

SOLiD



Sequencing by Oligo Ligation and Detection

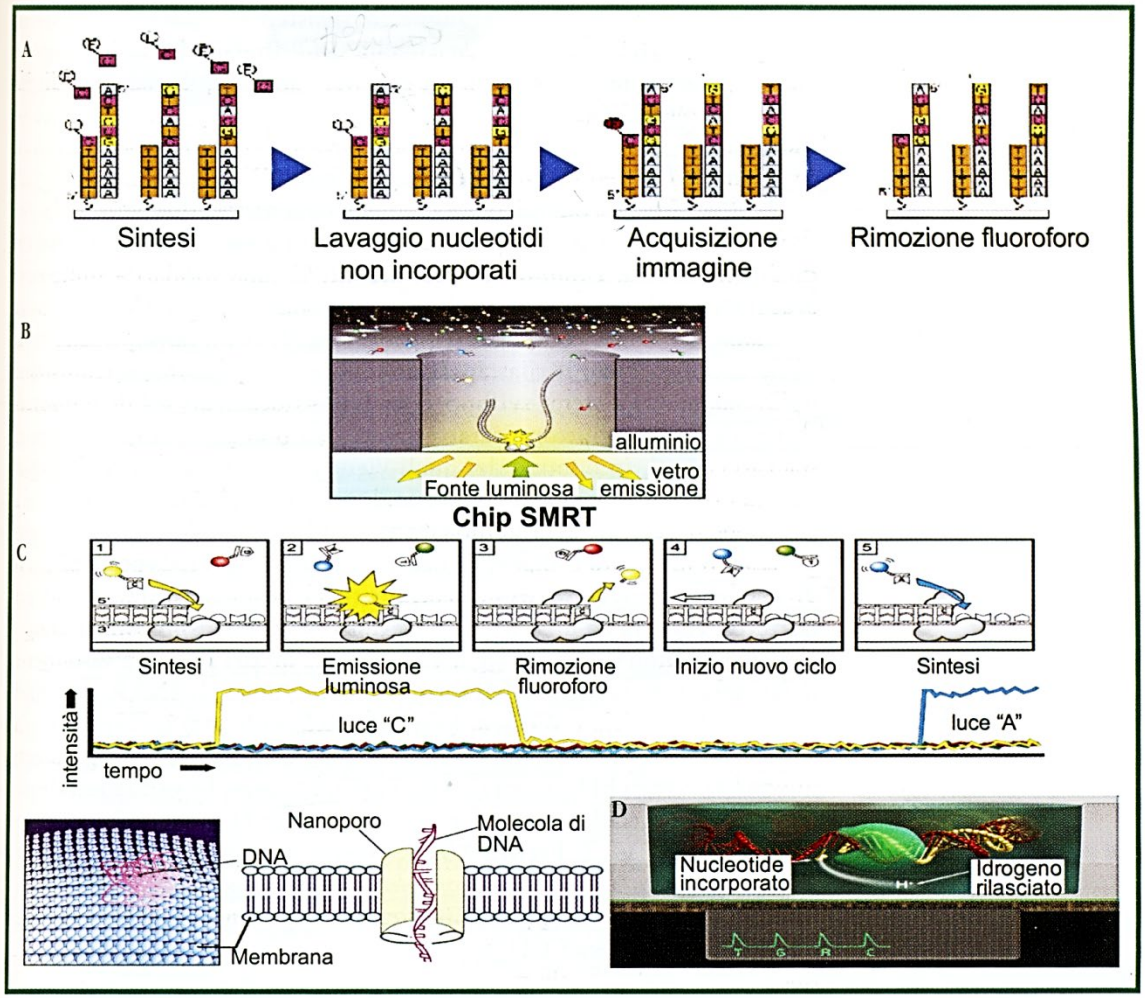
Ligazioni in successione di ottameri marcati con diversi fluorofori
Frammenti di DNA legati ad adattatori denaturati e fissati su nano sfere.
Amplificazione mediante PCR in emulsione
Le nanosfere sono singolarmente fissate su un vetrino per essere sequenziate
Primer n complementare agli adattatori
Si aggiungono gli ottameri di cui si conoscono solo le prime due basi
Un ottamero si lega a valle del primer ad opera della T4 DNA ligasi ed emette fluorescenza



Per riconoscere l'esatta sequenza è necessario conoscere il primer e le basi in posizione 0, -1 e -2
Ogni base è determinata due volte per questo la determinazione è affidabile

Terza generazione (next next generation sequencing, NNGS)

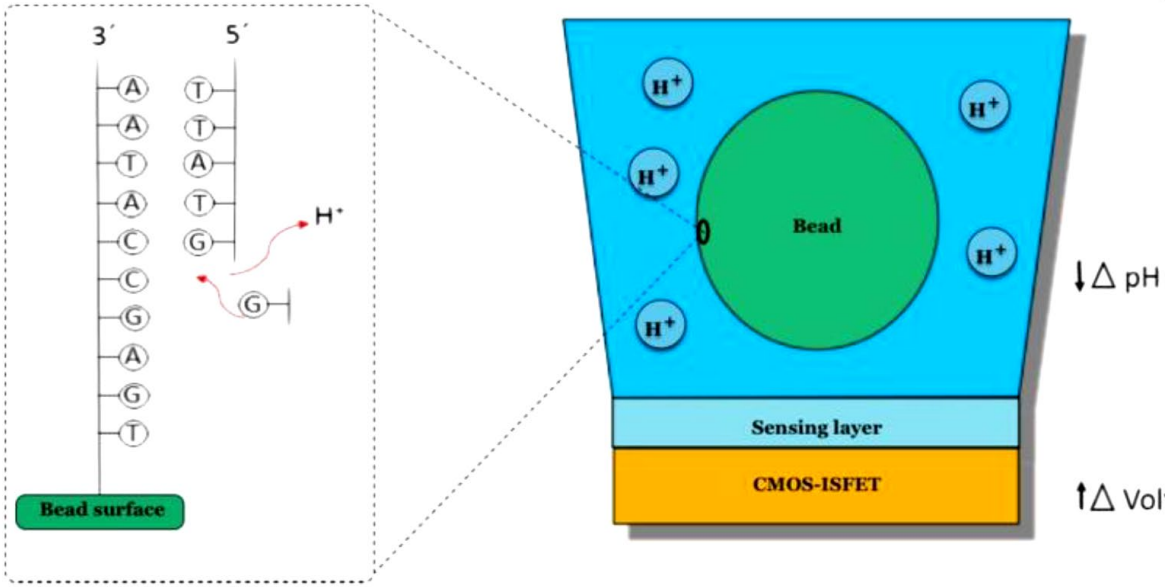
Sequenziamento diretto di singole molecole di DNA



-Limitata quantità di DNA necessaria
-Non richiede amplificazione della sequenza bersaglio
-Possibilità di ottenere molti dati per unità di tempo

-Sono però suscettibili di errori per l'estrema sensibilità degli strumenti

Ion Torrent



Nella piattaforma Ion Torrent, il chip è il sequencer.
Ogni pozzetto del chip agisce come un misuratore di pH che è in grado di rilevare i cambiamenti nella concentrazione di H+ prodotta nella polimerizzazione del DNA (Garrido-Cardenas J. A. et al., 2017)

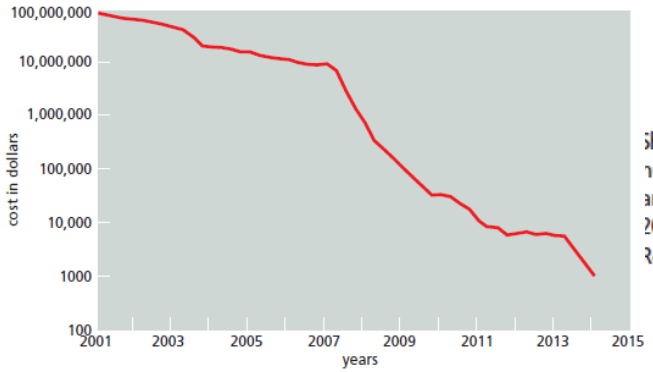
Supporto solido di materiale semiconduttore con elevato numero di pozzetti
Ogni pozzetto contiene un filamento di DNA
Ad ogni ciclo viene aggiunto un diverso nucleotide e se viene incorporato ad opera della polimerasi ci sarà rilascio di ione idrogeno e la variazione di pH sarà rilevata

I tempi di esecuzioni sono molto veloci infatti una corsa richiede solo poche ore ed è molto versatile. Per contro questa tecnologia ha, come il 454, un alto tasso di errore

ION TORRENT™ SEQUENCING

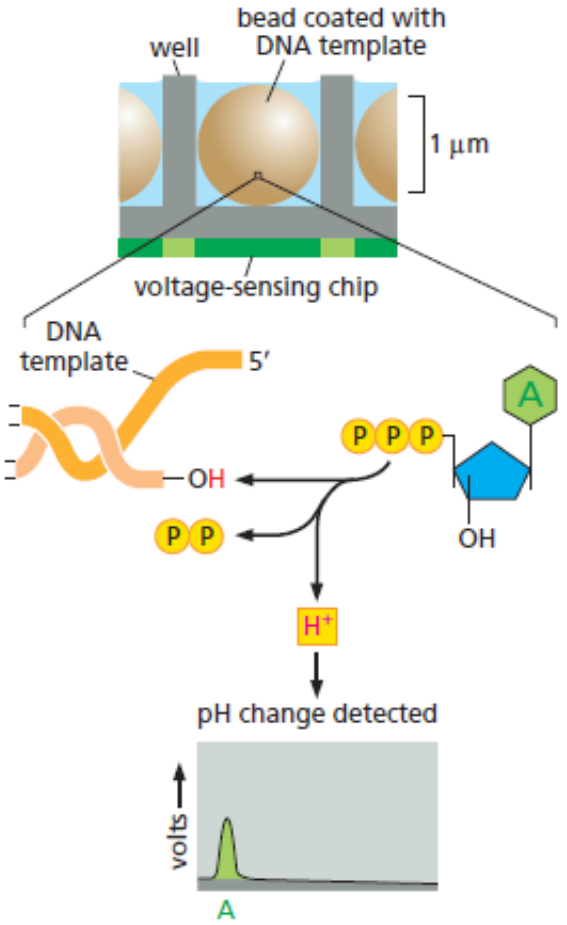
Another widely used strategy for rapid DNA sequencing is called the *ion torrent* method. Here, a genome is fragmented, and the individual fragments are attached to microscopic beads. Using PCR, each DNA fragment is then amplified so that copies of it eventually coat the bead to which it was initially attached. This process produces a library of billions of individual beads, each covered with identical copies of a particular DNA fragment.

Like eggs in a carton, the beads are placed into individual wells on an array that can hold a billion beads in a square inch. Beginning with a primer, DNA synthesis is then initiated on each bead. A hydrogen ion (H^+) is released (along with pyrophosphate) each time a nucleotide is incorporated into a growing DNA chain (see Figure 5–3), and the ion torrent method is based on this simple fact. Each of the four nucleotides is washed in, one at a time, over the array of beads; when a nucleotide is incorporated in the DNA of a given bead, the release of an H^+ ion changes the pH, which is registered by a semiconductor chip placed beneath the array of wells. In this way, the DNA sequence on a given bead can be read from the pattern of pH changes observed as nucleotides are washed over them. Like a high-resolution sensor in a digital camera, the ion torrent semiconductor chip can register enormous amounts of information and can thus keep track of billions of parallel sequencing reactions. Using this technology it is currently possible, using a single chip, to determine the nucleotide sequences of several human genomes in just a few hours.

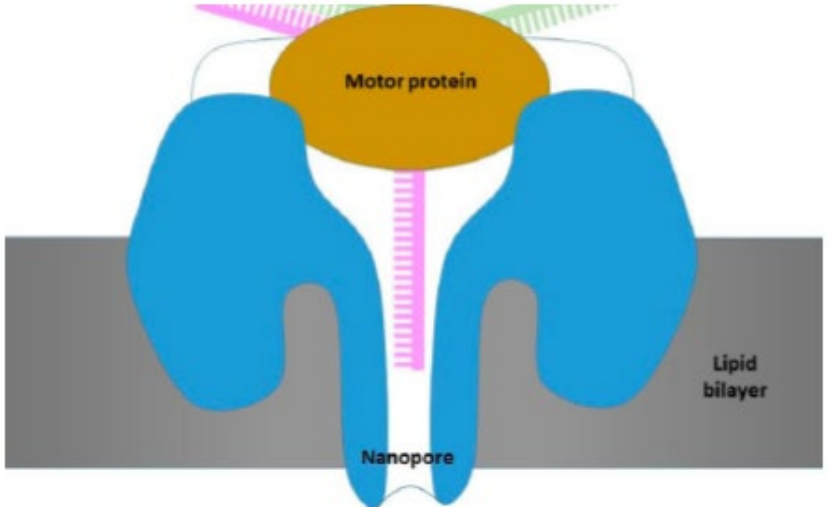


Shown here are the costs of sequencing a human genome, which was $100 million in 2001 and about a thousand dollars by the end of 2014. (Data from the National Human Genome Research Initiative.)

DNA sequencing by the ion torrent method. Beads, each coated with a DNA molecule that has been amplified many times, are placed in wells along with primers and DNA polymerase. As nucleotides are sequentially washed over the beads, those incorporated by the polymerase cause a pH change. In the example shown, an A is incorporated; thus, the template must have a T in this position. As the four nucleotides are sequentially washed over the beads, the sequence of the DNA on each bead can be "read" by the pattern of pH fluctuations. Billions of beads are monitored at once by a voltage-sensitive semiconductor chip placed below the array of beads.



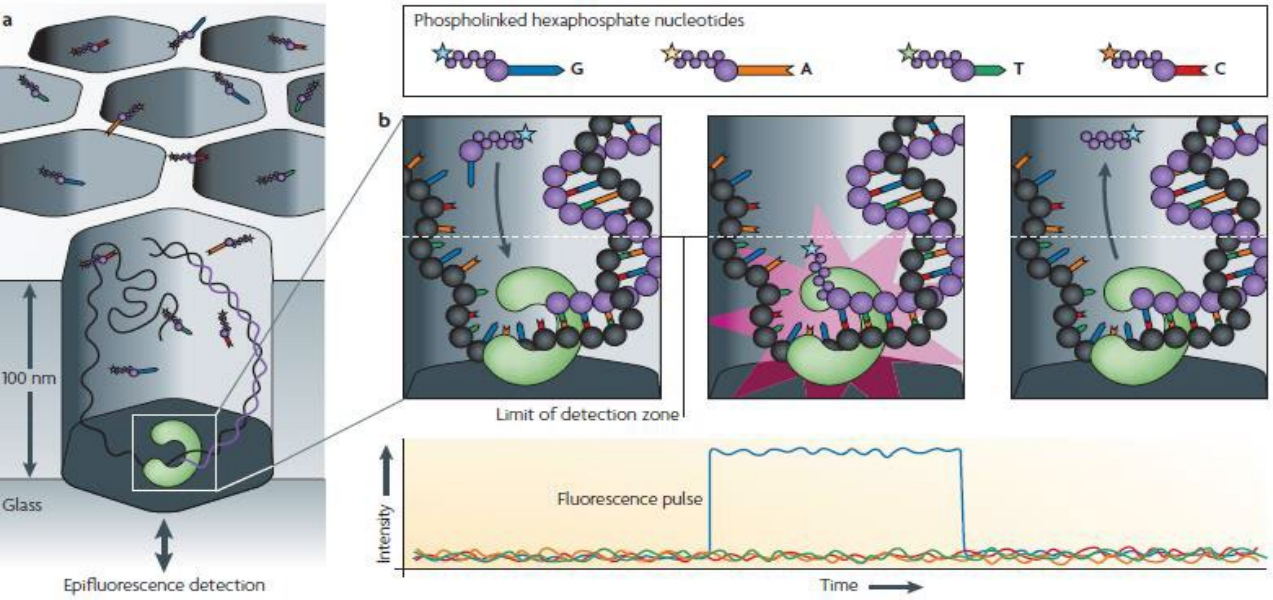
Nanopore



Canali utilizzati dalla piattaforma Nanopore Oxford per la sequenza del DNA. Il passaggio del DNA attraverso il nanopore produce alterazioni che vengono misurate grazie alle variazioni di tensione rilevate (Garrido-Cardenas J. A. et al., 2017)

La durata della traslocazione di un polinucleotide attraverso un nanoporo dipende dalla sua sequenza
Il sistema è in grado di sequenze che differiscono anche di un solo nucleotide

PacBio- SMRT (Single Molecule Real Time sequencing)



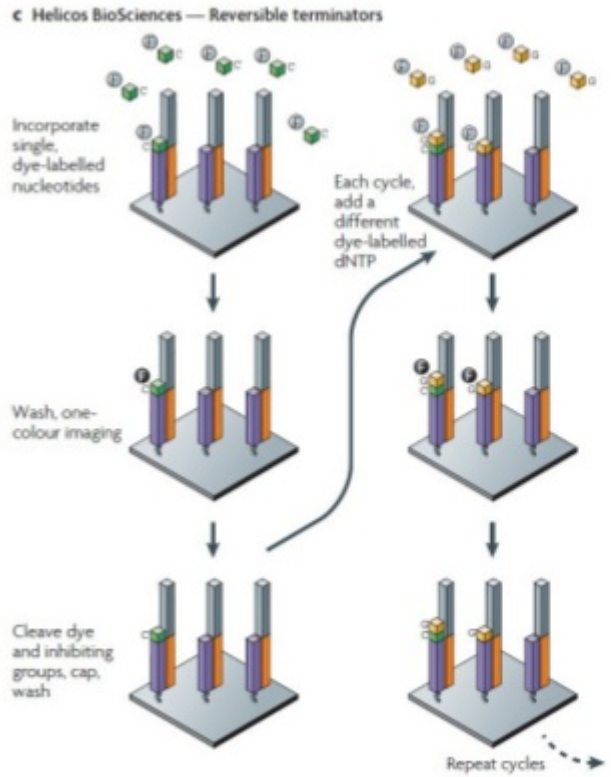
Chip costituito da film metallico contenente micropozzetti all'interno dei quali vengono inseriti filamenti di DNA
A contatto con il chip c'è una superficie trasparente al di sotto della quale è presente una fotocamera a rivelazione di fluorescenza

Nucleotidi marcati con 4 fluorofori diversi- quando un dntp si incorpora mediante DNA polimerasi il fluoroforo viene rilasciato ed emette fluorescenza che viene rilevata dalla fotocamera

Helicos

Helicos BioSciences

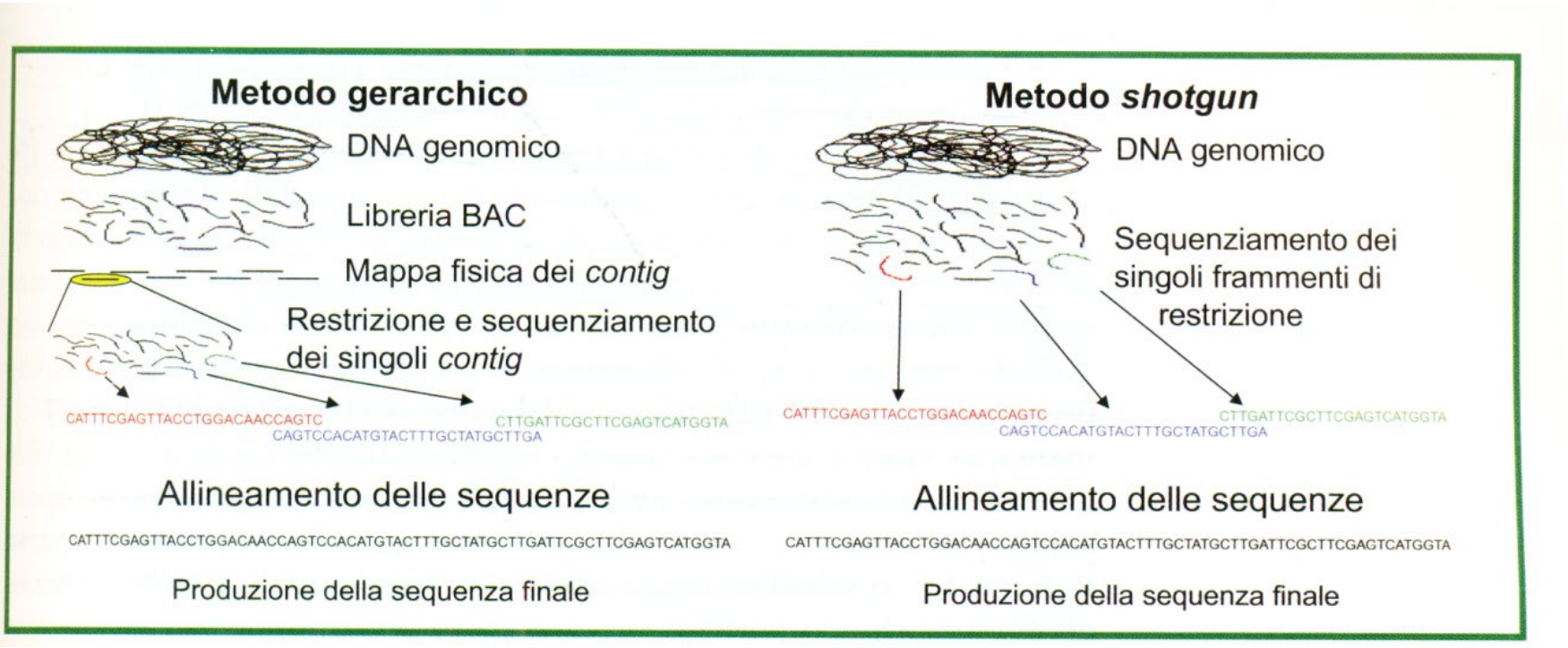
2. Sequencing



- Each cycle consists of:
  1. adding the polymerase and **one** of the labeled nucleotide
  2. rinsing, imaging of multiple positions
  3. cleavage of the dye labels
- 224 cycles were performed to sequence the genome of the M13 virus to an average depth of >150X with 100% coverage

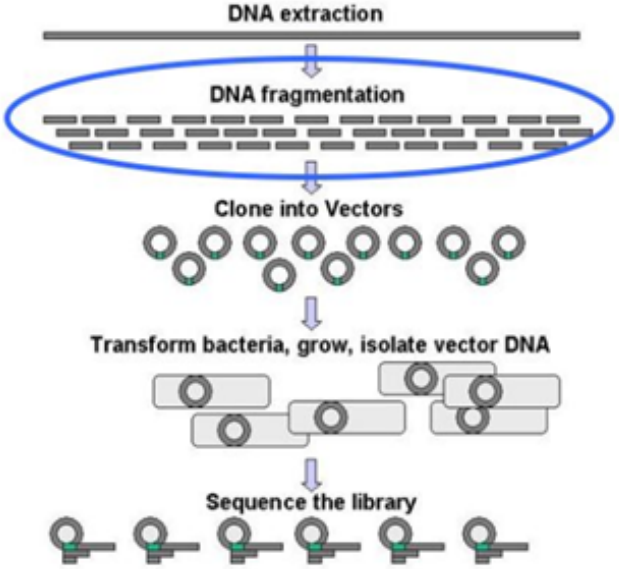
Sequenziamento di molecole di DNA senza fase di amplificazione
Si utilizza una libreria di frammenti ognuno legato ad un poli A
I poli A sono ibridati a poli T
La sequenza è determinata aggiungendo uno per volta nucleotidi marcati

Sequenziamento di un genoma
Due strategie di frammentazione e ricostruzione dei genomi



CLONAGGIO E SEQUENZIAMENTO

Per sequenziare il DNA è importante che i frammenti di sequenza ignota siano clonati in appositi vettori. I frammenti sono ottenuti mediante frammentazione del genoma o, nel caso di produzione di EST (Expressed Sequence Tags), mediante retrotrascrizione dell'mRNA.



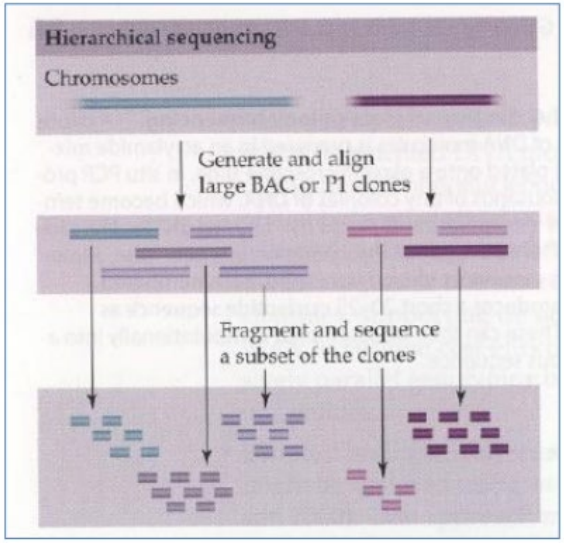
Per sequenziare il frammento inserito vengono utilizzati come *primer* (oligonucleotidi) di innesco delle sequenze del vettore stesso. I primer in questo caso prendono il nome di **primer universali**. In genere i primer sono progettati sia in posizione 5' (**Forward** primer - For) rispetto all'inserto, che in posizione 3' (**Reverse** primer - Rev).

Per il sequenziamento Sanger, la lunghezza media di buone sequenze si aggira attorno a 500–800 nucleotidi, perchè la polimerasi in genere perde la sua attività dopo avere incorporato un migliaio di nucleotidi. E' per questo che in posizione 3' della sequenza non si hanno buone incorporazioni dei nucleotidi modificati causando una perdita di bontà della sequenza incognita.

Oltre a problematiche nelle estremità della sequenza incognita, particolari problemi di sequenziamento possono essere dati dalle sequenza ripetute. Queste causano uno slittamento della polimerasi e quindi una sorta di perdita del frame di polimerizzazione.

Metodo gerarchico

Produzione di frammenti da 100 a 300 kb
Clonaggio in vettori BAC e trasferimento dei vettori in E.Coli
I cloni BAC sono organizzati in una serie ordinata e sovrapponibile di frammenti *contig*
Devono avere il minor grado di sovrapposizione ed equivalenti ad una regione cromosomica (minimal tilling path) che si ottiene mediante BAC fingerprinting oppure tramite Chromosome walking



BAC fingerprinting
Identificazione di sequenze comuni a più cloni mediante confronto dei profili di restrizione del DNA

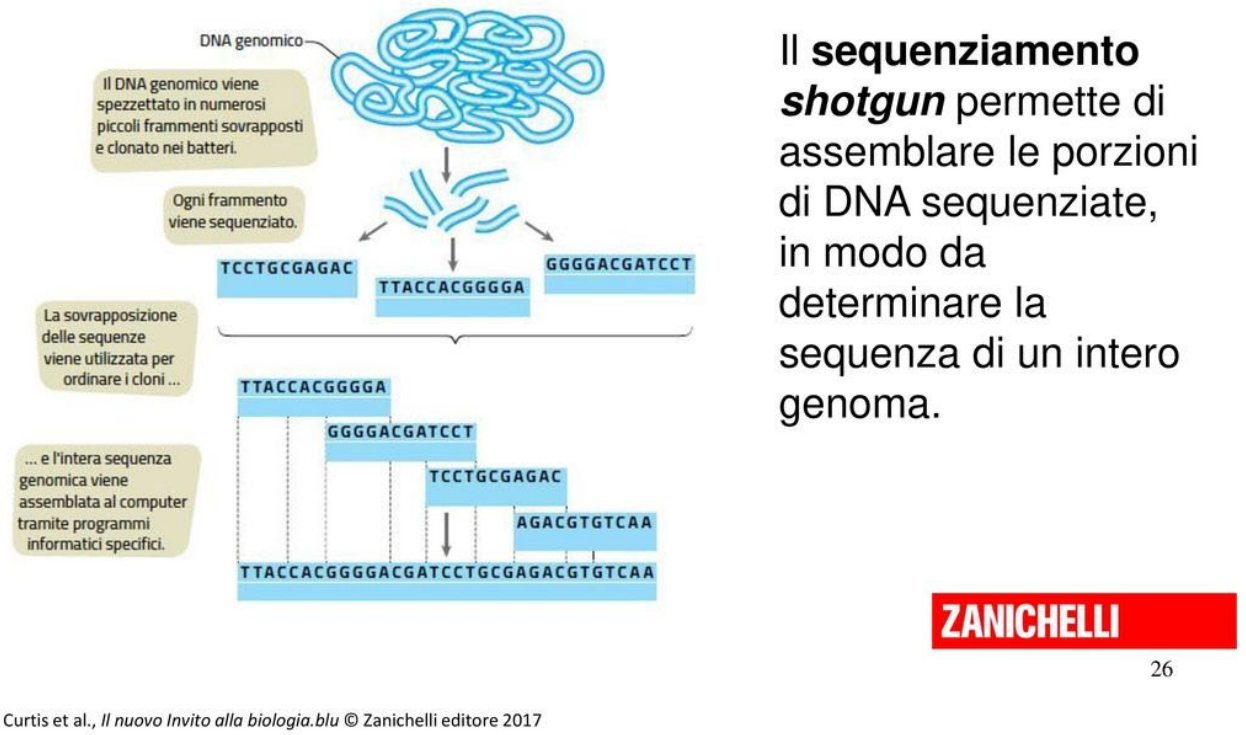
Chromosome walking
Identificazione di sequenze comuni a più cloni mediante ibridazione con una sonda marcata di piccole dimensioni

Una volta stabilito l'ordine dei cloni, questi vengono frammentati e di nuovo clonati e sequenziati
Assemblando in successione i singoli cloni BAC si ottiene la sequenza dell'intero genoma

Metodo shot gun

Sequenziamento di una libreria di frammenti di DNA genomico
Ogni base del campiona deve essere sequenziata più volte per garantire un'affidabile identificazione delle basi e perché i frammenti non sono distribuiti in modo uniforme

Copertura: È possibile che molti nucleotidi siano rappresentati in pochi frammenti ed altri da molti frammenti. Questo parametro indica la ridondanza del sequenziamento



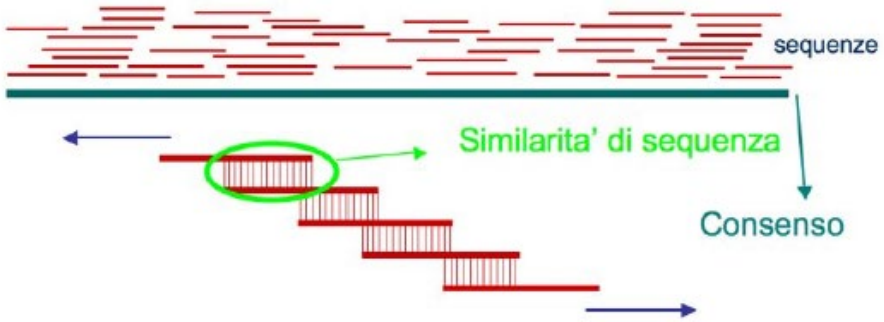
Curtis et al., *Il nuovo Invito alla biologia.blu* © Zanichelli editore 2017

Copertura C= NL/G
N=numero di frammenti
L= lunghezza media
G= lunghezza del genoma (bp)

Assemblaggio dei frammenti sequenziati

Nell'assemblaggio di sequenze si generano dei contigui (**contig**) per poi definire un consenso (**consensus**). Al fine di produrre un consenso è necessario che ogni posizione (base) sia validata.

Per avere una validazione delle singole posizioni si producono numerose sequenze della stessa porzione genomica generando una copertura maggiore di 1X:

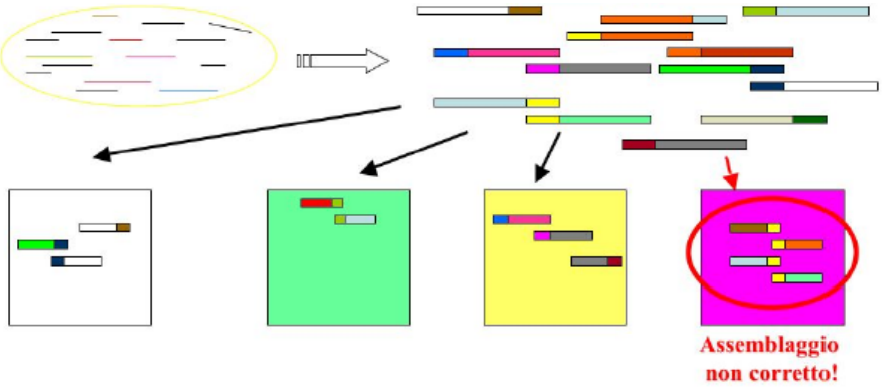


Con il termine copertura (**coverage**) vengono indicate quante basi in più sono sequenziate in relazione alla lunghezza totale di un genoma. Ad esempio una copertura di 1X sta ad indicare che il numero di basi sequenziate è uguale alla lunghezza totale del genoma incognito, ma questo non vuol dire che tutto il genoma viene sequenziato. Infatti, dopo il clonaggio e la trasformazione batterica, i cloni da sequenziare vengono recuperati con un processo casuale che implica la perdita di alcuni frammenti ed il recupero multiplo di cloni con il medesimo inserto. In relazione al recupero casuale dei cloni da sequenziare è possibile stimare mediante una distribuzione di Poisson la quantità di genoma mancante in funzione della copertura di sequenziamento effettuata. La funzione di Poisson è descritta dalla seguente equazione: $P_0=e^{-m}$ dove m rappresenta il fattore di copertura e P_0 la probabilità che una base non venga sequenziata.

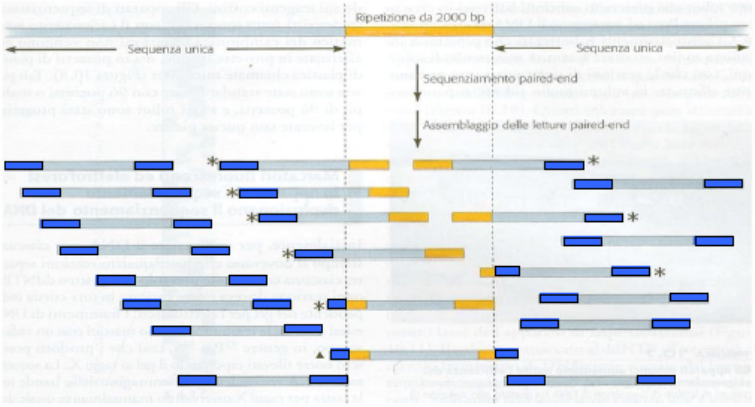
Con copertura 1X il 37% del genoma risulta non sequenziato. Oltre al recupero casuale dei cloni da sequenziare, bisogna considerare anche il fatto che alcune sequenze sono più difficili da clonare di altre. E' perciò necessario sequenziare un numero di basi pari a diverse volte la sequenza completa, per arrivare a trovare un numero di sovrapposizioni sufficiente. La copertura di una porzione di genoma non si basa solamente sul numero di volte che viene sequenziata la specifica porzione, ma anche sulla qualità di sequenze prodotte.

Problemi di assemblaggio

Il sequenziamento di entrambe le estremità dei cloni, senza doversi preoccupare di avere una sequenza completa dell'inserto è di fondamentale importanza anche per l'assemblaggio di regioni genomiche con **sequenze ripetute**. Ciò tuttavia può creare problemi, quando una sequenza ripetuta è presente all'estremità di frammenti diversi:



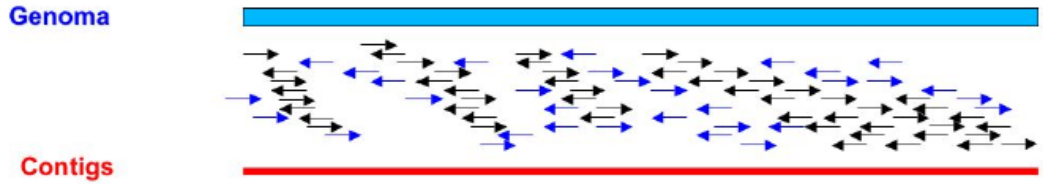
Nella figura che segue, le parti grigio-celesti rappresentano le regioni non sequenziate. Le sequenze derivanti dai cloni con asterisco permettono l'assemblaggio della regione con ripetizioni (in giallo), perché a queste sono associate sequenze non ripetute (in blu). Il clone che copre la sequenza ripetuta (marcato da un triangolo) permette di confermare l'assemblaggio della sequenza ripetuta:



Va ribadita l'importanza di sequenziare entrambe le terminazioni dell'inserto: in assenza di cloni che fanno da ponte, le sequenze ripetute non sono risolvibili.

Sequenziamento shotgun (WGS)

L'assemblaggio delle sequenze risulta più semplice se si utilizzano cloni contigui e sovrapposti di elevate dimensioni, ma questo approccio richiede tempo per il finishing e la copertura dei gap. L'approccio di sequenziamento shotgun (***Whole Genome Shotgun*** o **WGS**) si basa sul clonaggio di frammenti multipli di piccole dimensioni per poi ricostruire la sequenza di contigui definitiva. Nel 1995, l'approccio shotgun è stato utilizzato per il sequenziamento del genoma di *Haemophilus influenzae* (1830 kb). La strategia ha previsto l'esecuzione di 28643 esperimenti di sequenziamento; quelli andati a buon fine hanno coperto 11631 bp con una ridondanza di circa 6.
L'approccio WGS può essere molto efficace per il sequenziamento di genomi piccoli, come ad esempio un genoma virale; esso prevede alcune fasi fondamentali:
- si crea una libreria di corti frammenti di DNA (1500 - 2000 bp) in vettori plasmidici
- si producono le sequenze dei frammenti corti: sequenze che coprono la stessa regione di DNA formano un contig ininterrotto
- aumentando il numero di sequenze molti contig si fondono tra loro
- producendo moltissime sequenze si ricostruisce l'intera sequenza del genoma

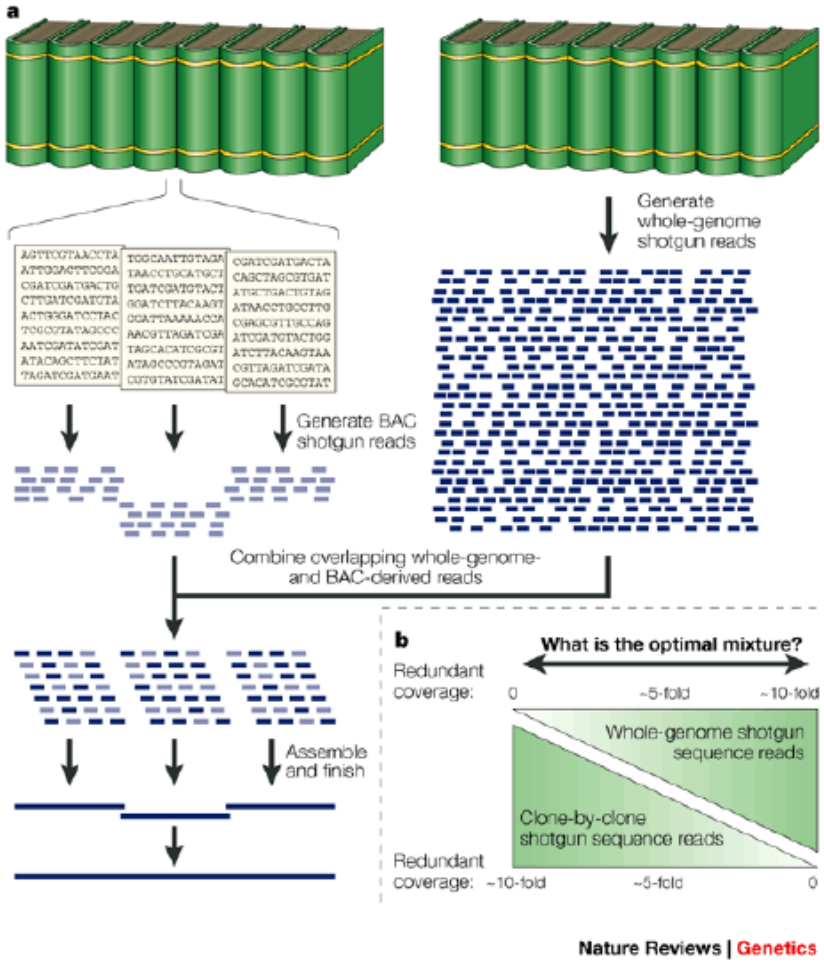


L'approccio shotgun a genomi di grandi dimensioni pone diversi problemi di assemblaggio:
- al crescere del numero dei frammenti aumenta enormemente il numero di overlap possibili, creando notevoli problemi computazionali;

- la presenza di regioni ripetute può determinare errori di assemblaggio con perdita di sequenze o unione erronea di frammenti, appartenenti anche a cromosomi diversi;
- il numero di gap finali da chiudere diviene molto alto e non gestibile facilmente con metodiche sperimentali

Pertanto, il metodo shotgun non è adeguato al sequenziamento completo di genomi complessi. Celera Genomics ha potuto utilizzarlo per il genoma umano solo perchè poteva accedere alla mappa fisica del consorzio pubblico. Per i genomi complessi si seguono strategie combinate (**hybrid approach**): si genera una mappa da utilizzare come canovaccio ed il sequenziamento WGS è eseguito sulle varie regioni del genoma, poi assemblate facendo riferimento alla mappa.

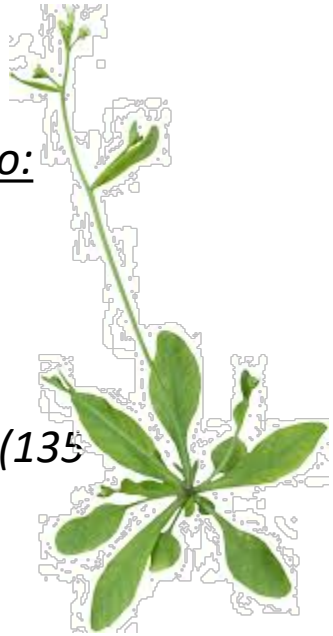
Nella figura che segue: si prepara una library di subcloni e numerose *sequence reads* (blu scuro) sono ottenute. Nel frattempo, anche BAC mappati individualmente sono soggetti a sequenziamento WGS. Le *sequence reads* derivanti dai BAC (blu chiaro) sono utilizzate per identificare le sequenze sovrapposte nella più ampia collezione di *sequence reads* ottenute, riducendo la complessità dei dati WGS. Il set combinato di *sequence reads* per ciascun BAC è infine assemblato e soggetto a finishing. Un coverage 8-10x è in genere necessario, ma il balance ottimale tra *sequence reads* ottenute clone-by-clone o WGS è variabile:



Il sequenziamento dei genomi di specie coltivate consente di scoprire geni e forme alleliche nuove la cui caratterizzazione consente di migliorare le piante utilizzate come fonte di cibo, fibre, energie rinnovabili ecc.

Esempi di genomi vegetali sequenziati: ***Arabidopsis thaliana***

Primo genoma vegetale sequenziato (2000)
<u>Considerato organismo modello:</u>
-breve ciclo vitale,
- taglia ridotta,
-elevato numero di semi per pianta,
-genoma di ridotte dimensioni (135 000 000 bp),
-si trasforma facilmente,
-numerose linee mutanti,
-mappe dettagliate genetiche e fisiche dei 5 cromosomi



120 Mbp sequenziati
30Mbp restanti sono sequenze ripetute o regioni difficili da sequenziare (tipo centromeri o regioni codificanti per RNA ribosomiali
27000 loci genici
1/5 dei geni subisce splicing alternativo
35000 trascritti diversi che codificano per proteine
5000 pseudogeni e elementi trasponibili
1300RNA non codificanti tra cui 177 miRNA
Densità genica pari a 4,35 kb/gene con circa 6 esoni per gene (lunghezza media 296 bp)

Il genoma plastidiale: 88 geni codificanti proteine, 37 codificanti pre t-RNA

Il genoma mitocondriale: 122 geni codificanti proteine, 21 codificanti pre t-RNA

Annotazione funzionale per il 60% dei geni il resto ha funzione sconosciuta

Esempi di genomi vegetali sequenziati: ***Oryza sativa***

Prima pianta monocotiledone sequenziata
Nel 2002 sequenziamento della ssp Indica
466 Mbp
È tra le più coltivate in Cina

Nel 2005 sequenziamento della ssp Japonica
372 Mbp
12 cromosomi
55986 loci di cui 17000 elementi trasponibili
39000 geni codifianti proteine
4,9 esoni per gene di lunghezza media di 318 pb
49000 trascritti diversi

Esempi di genomi vegetali sequenziati: ***Solanum lycopersicum***

Sequenziamento iniziato nel 2004 e terminato nel 2012 da un consorzio di ricercatori provenienti da 14 nazioni

Genoma costituito da 950 Mpb
Ripartito per 12 cromosomi il cui 75% è eterocromatina
La maggioranza dei geni è localizzata nelle porzioni distali dei bracci cromosomici in lunghi tratti eucromatici

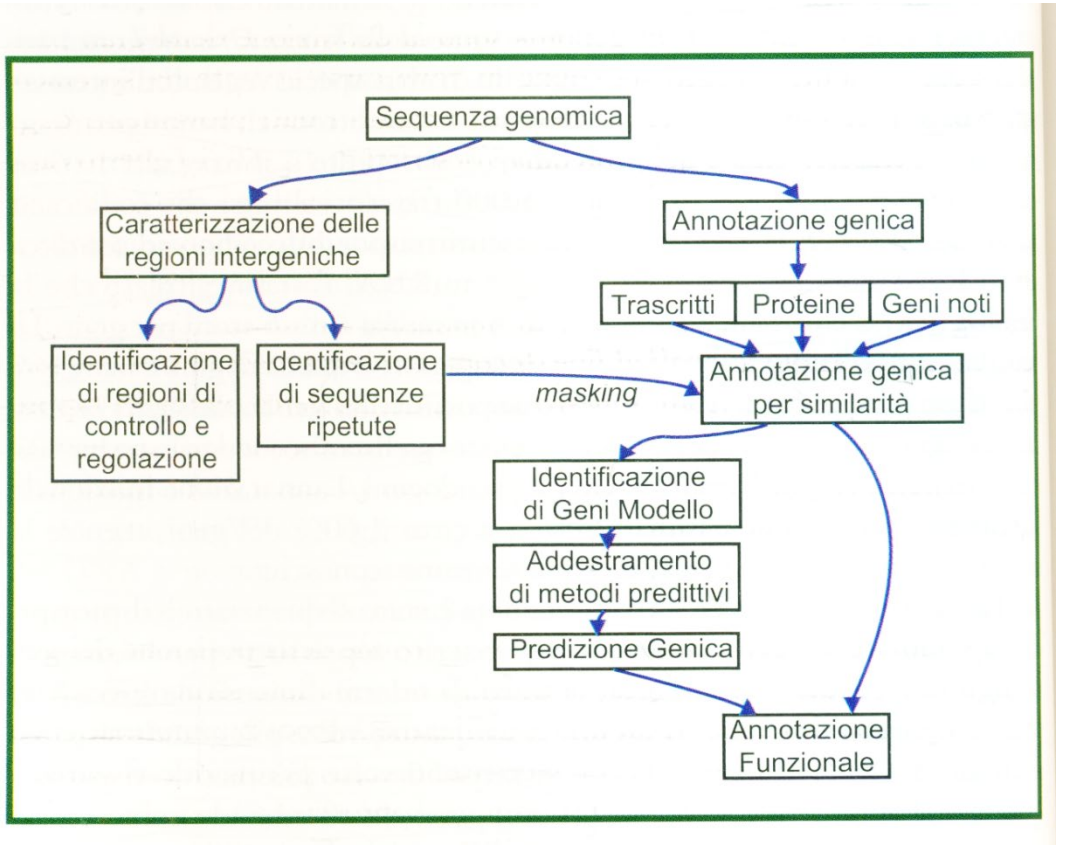
Annotazione dei genomi

Descrizione e caratterizzazione delle sequenze
Descrizione delle regioni geniche codificanti mRNA, tRNA, rRNA
Caratterizzazione delle regioni non geniche con particolari proprietà strutturali: sequenze ripetute, regioni telomeriche, centromeriche e pericentromeriche, regioni di controllo e di legame con specifiche proteine

Tutto ciò si può ottenere attraverso l'uso di metodologie bioinformatiche

Bioinformatica: scienza multidisciplinare che analizza l'informazione biologica con metodi computazionali al fine di comprendere i meccanismi che sono alla base dei processi vitali

L'annotazione dei geni prevede:
-L'identificazione della posizione dei geni lungo la sequenza genomica
-La definizione della loro struttura
-La descrizione della loro funzione



Annotazione dei geni

Basata sulla ricerca di similarità

Attraverso…
-Confronto della sequenza del genoma con sequenze ottenute da esperimenti di trascrittomica o proteomica dell'organismo in esame
Oppure
-confronto con geni,trascritti o proteine identificati e caratterizzati in organismi filogeneticamente correlati

Consente anche di identificare dei geni che vengono poi presi come modello ed utilizzati per l'annotazione con il metodo predittivo

Software: BLAT, SIM4,Genome Threader

Basata su metodi predittivi

Attraverso…
-utilizzo dei geni modello per addestramento dei metodi predittivi *ab initio*
Si identificano quindi regioni promotrici, segnali di regolazione, contenuto in GC, siti di splicing, di inizio e di fine ecc.
Si procede poi con l'identificazione dei geni sull'intera sequenza del genoma

Annotazione dei geni non codificanti mRNA (non-coding RNA)

rRNA, tRNA, microRNA

Per l'annotazione di queste sequenze è necessario non solo il confronto con altri RNA noti e quindi la struttura primaria ma anche la struttura secondaria a trifoglio per il tRNA e a forcina per i microRNA

Annotazione della porzione non codificante del genoma Regioni di sequenza in grado di legare fattori trascrizionali e proteine regolatrici (database TRANSFAC)

Regioni contenenti sequenze ripetute (telomeri, centromeri, elementi trasponibili ecc.)

La loro identificazione è importante

- per analisi di tipo evolutivo e comparativo
- Per evitare che queste regioni (spesso in parte simili a sequenze geniche) inficino l'annotazione di sequenze geniche. Per questo esistono programmi che ne effettuano il mascheramento

Risulta quindi importante identificare de novo famiglie di elementi ripetuti per costruire librerie di riferimento accurate

Software: RECON e Repeat Scout

Annotazione funzionale

Al fine dell'annotazione è necessario caratterizzare anche dal punto di vista funzionale

-si ottiene se si conosce la regione codificante e la funzione della proteina stessa

-Se non si conosce allora si può ipotizzare mediante ricerca di similarità con prot. Note oppure si indentifica come proteina con funzione sconosciuta

Software: GenPept, UniProt per identificazione di similarità. PFAM SMART PRINTs BLOCKs per info su domini proteici, motivi proteici, regioni conservate.

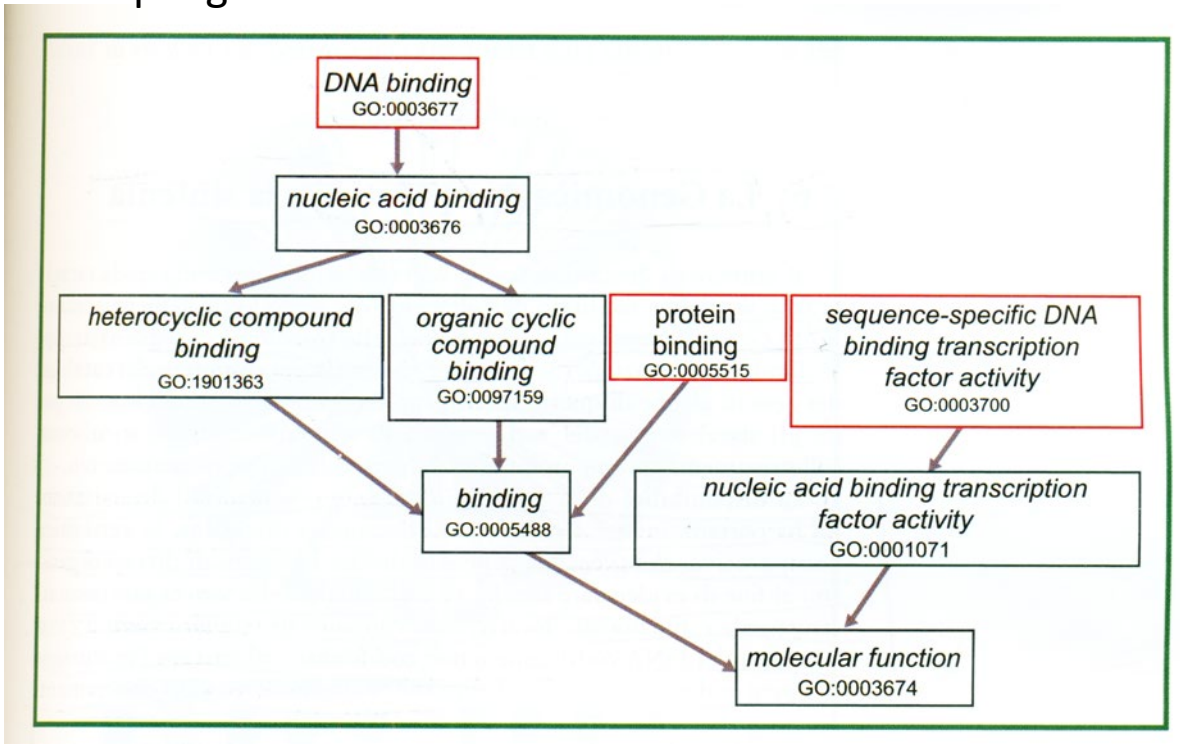
GENE ONTOLOGY (GO)

Obiettivi: uniformare la descrizione delle funzioni geniche per costruire un vocabolario ben definito per essere utilizzato nell'annotazione delle sequenze

Obiettivi: unificare la descrizione delle molecole rappresentate dai dati contenuti in differenti banche dati biologiche

Esempio gene MYB AT3G27920 *di A. thaliana*

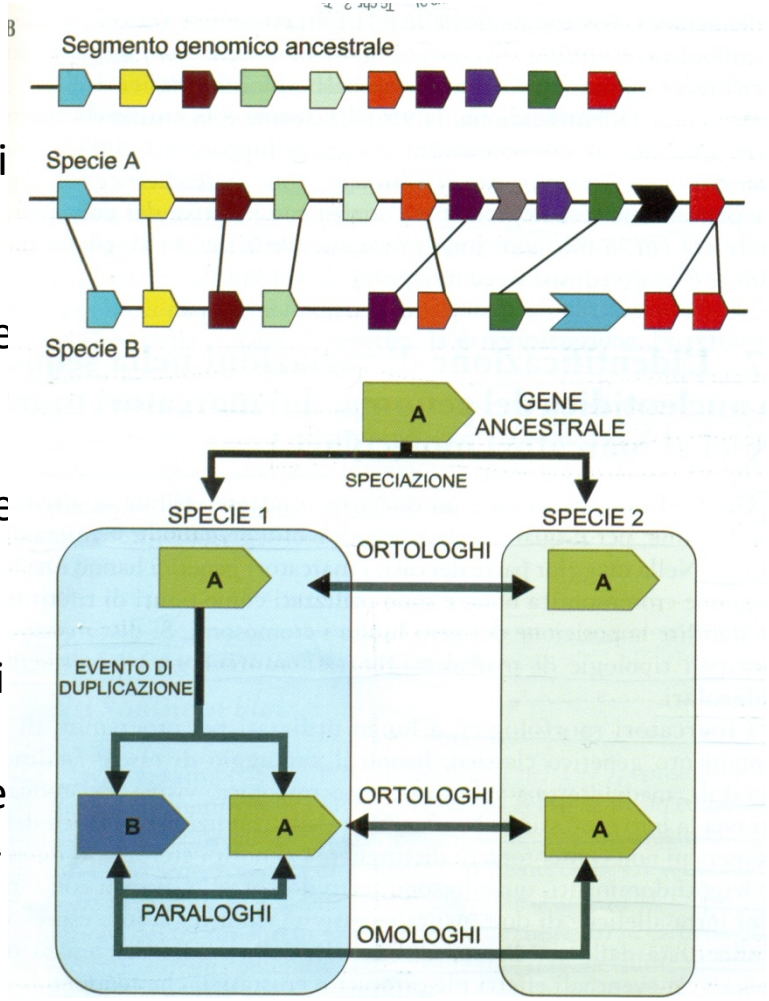
Tre domini distinti
-componente cellulare
-processi biologici
-funzione molecolare



Genomica comparativa e sintenia

Consiste nel confrontare genomi di diversi organismi per evidenziare similarità e differenze nella loro organizzazione strutturale e funzionale

Consente inoltre di studiare come i cambiamenti del DNA codificante e non, influenzino l'evoluzione dei geni e determinino divergenze nell'espressione e negli adattamenti specie-specifici



Regioni cromosomiche di specie diverse o di regioni cromosomiche duplicate presenti nella stessa specie che includono un insieme di geni che condividono lo stesso ordine relativo (co-linearità)

Obiettivo della genomica comparativa è la ricerca di SINTENIE

Geni **omologhi**:geni che condividono una comune origine evolutiva in uno stesso genoma o tra genomi. Si distinguono in
ortologhi (origine a seguito di un evento di speciazione) e
paraloghi (a seguito di un evento di duplicazione)

Sintenia e colinearità

I genomi degli eucarioti differiscono

nel grado in cui i geni rimangono sullo stesso cromosoma

↓

sintenia

nel grado in cui l'ordine dei geni viene mantenuto sul cromosoma

↓

colinearità

Nel genoma delle angiosperme c'è grande differenza nelle dimensioni e arrangiamento genico (anche tra specie vicine)

Duplicazioni di interi genomi
Perdita di intere regioni cromosomiche

Dimensione dei genomi variabile di 1000 volte

N. di cromosomi variabile di 50 volte

Due organismi con un antenato in comune relativamente recente hanno genomi che presentano differenze specie-specifiche, basate sullo schema comune del genoma ancestrale

Quanto più due organismi sono vicini nella scala evolutiva, tanto più correlati sono i loro genomi

sintenia

la parziale o completa conservazione dei geni sul cromosoma

è possibile utilizzare le informazioni di mappa di un genoma per localizzare geni nel secondo genoma

Nelle piante la sintenia è di enorme importanza può risultare utile per gli studi di genomica comparata e operazioni di mappatura

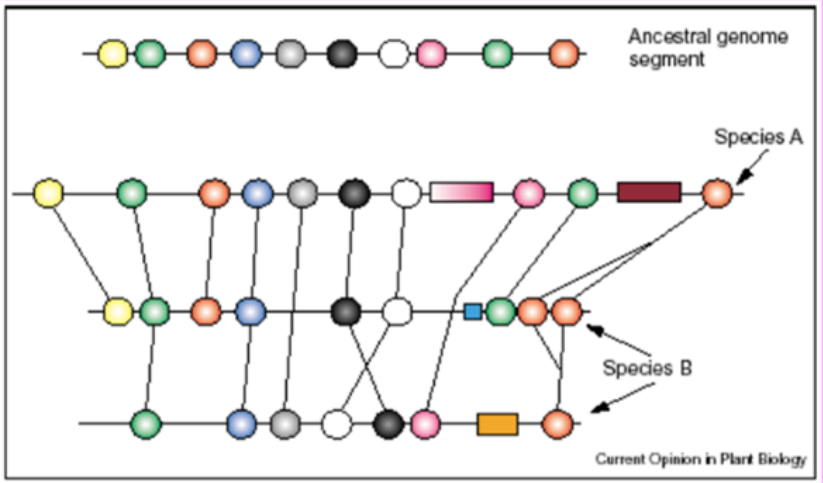
Il frumento ha un genoma molto grande (17.000 Mb) più di cinque volte quello umano

Una pianta modello con un genoma più piccolo ma sintenico sarebbe utile…

Il genoma di riso è di 400 Mb e la genomica comparativa tra i due ha rivelato molte similarità

mappando inizialmente la posizione del gene equivalente su un genoma più piccolo di riso e per sintenia posizionarli su un genoma più grande

La sintenia nelle piante



I geni mantengono lo stesso ordine e spesso anche l'orientamento

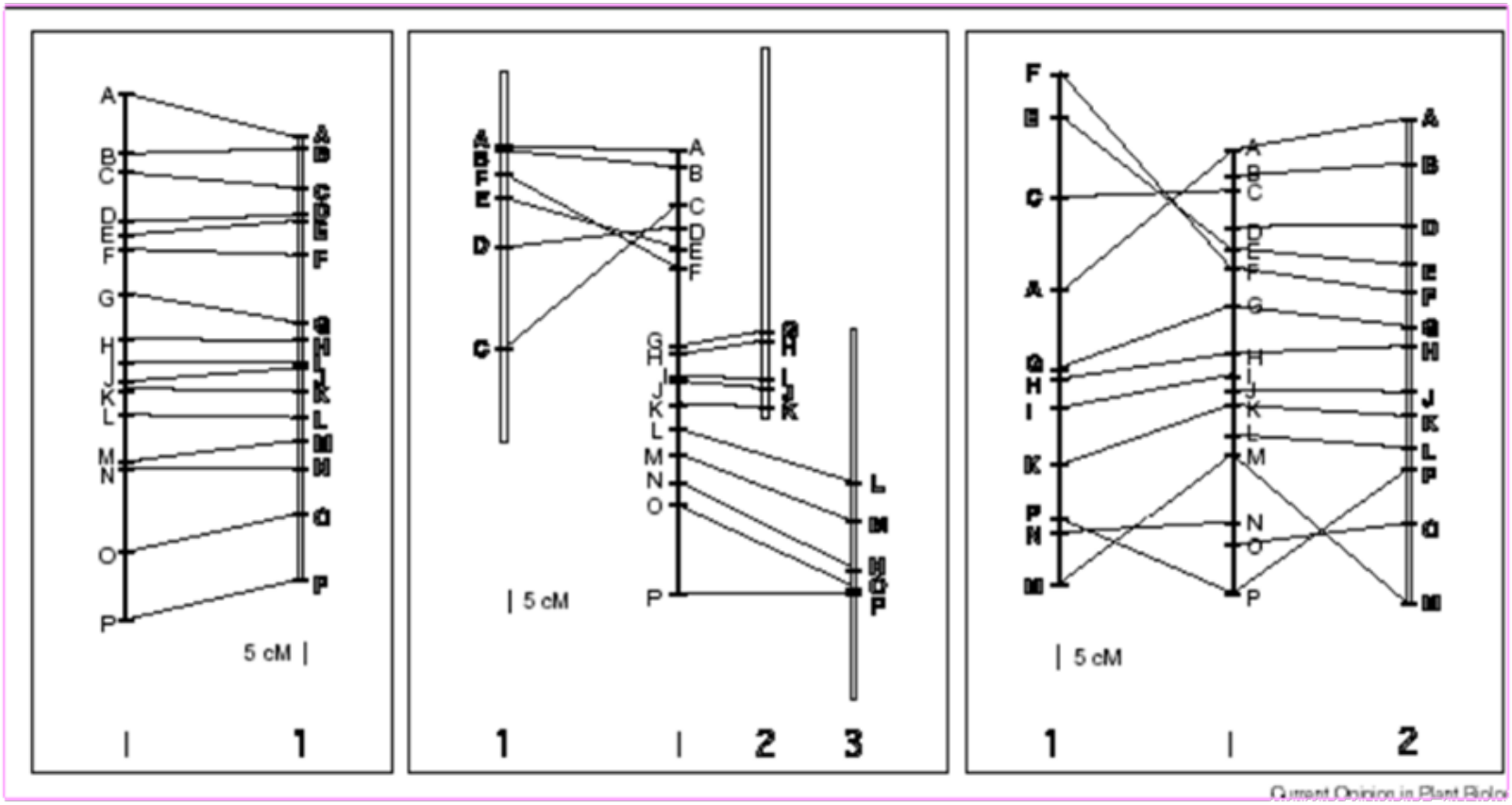
1. Sintenia importante in genomica comparativa, rivela l'evoluzione dei genomi di specie correlate

2. Condividere la sintenia di un frammento genomico, significa avere un ancestrale comune

3. Geni sintenici sono ortologhi localizz in un frammento sintenico. Funzione?

4. Studi di sintenia importanti per le funzioni geniche e per l'evoluzione del genoma

Comparing *Arabidopsis* to other flowering plants



Mappe lineari

Colinearità con diversi segmenti cromosomici

traslocazioni e inversioni

Confronto tra diploide e tetraploide

mancano B e N

Pattern di microsintenia

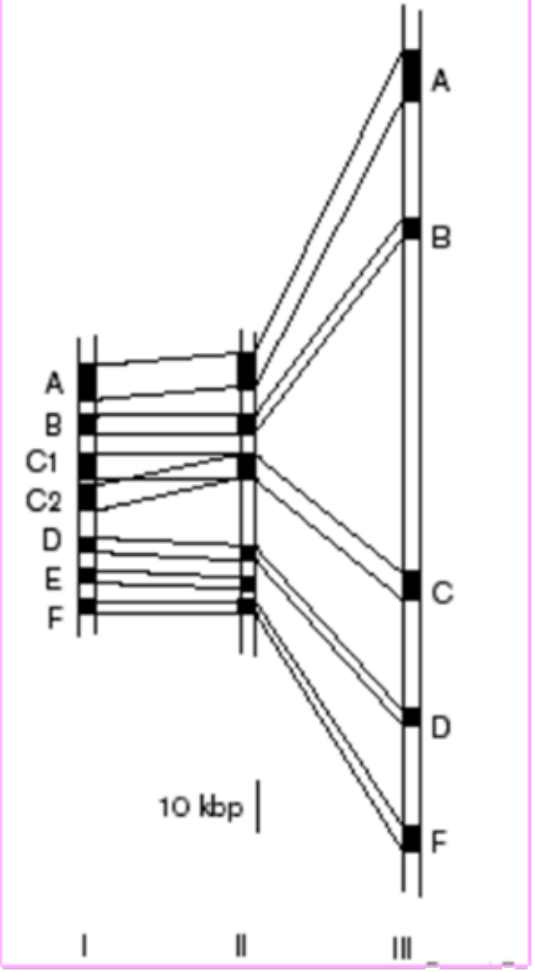
Un'analisi comparativa delle regioni genomiche ortologhe derivate da diverse specie (I-III)

un'elevata conservazione di sequenze geniche

le sequenze intergeniche non mostrano omologie significative

La microsintenia ha messo in evidenza delezioni e duplicazioni di sequenze geniche

Il gene E è deleto nella specie III, mentre la specie I porta due copie del gene C



La comparazione dell'arrangiamento di geni in specie con diverse dimensioni del genoma (II e III), mostra che alcune delle differenze in termini di dimensioni del genoma possono essere attribuite alle diverse dimensione delle regioni intergeniche

I genomi delle graminacee

Species	Genome Size	Approx. Gene #	Predicted kb/gene
Barley	4800 Mb	30,000	160
Maize	2500 Mb	50,000	50
Rice	430 Mb	30,000	15
Sorghum	750 Mb	30-50,000	15-25
Diploid Wheat	5300 Mb	30,000	175

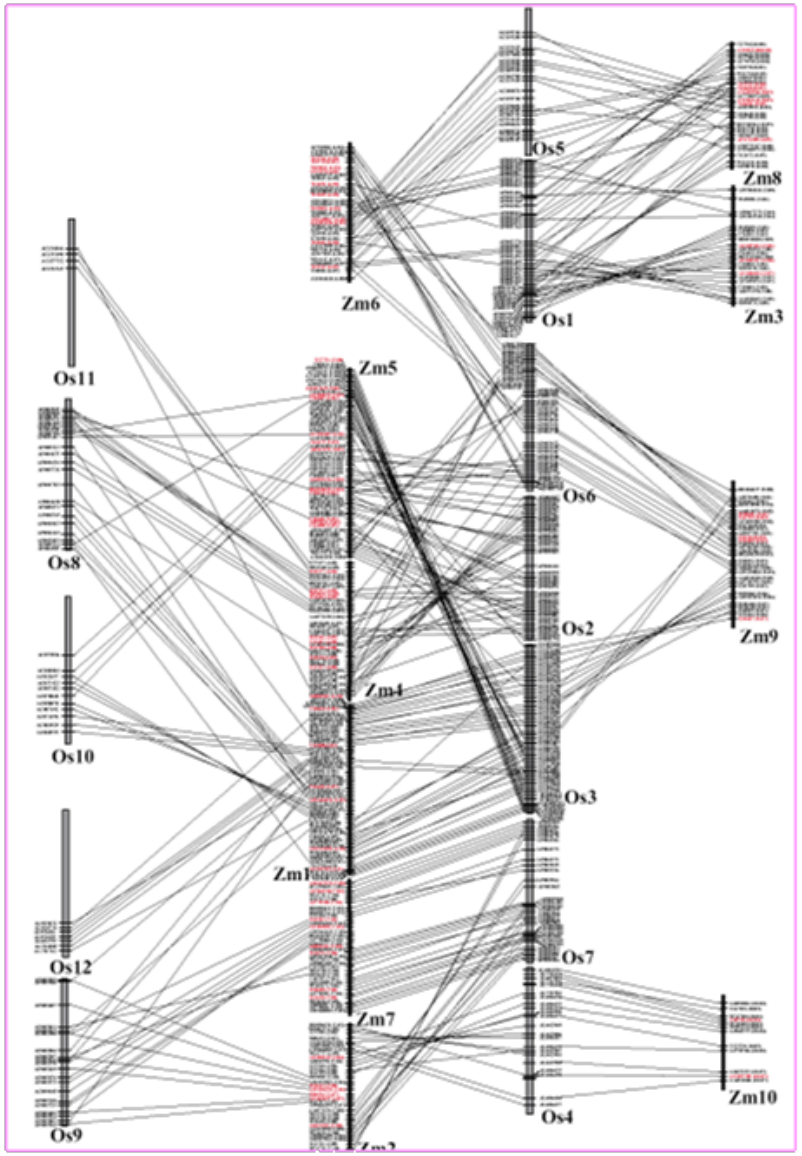
- **Enormi differenze in termini di dimensioni del genoma dei cereali!**
- **sono rinvenibili estese regioni di sintenia**
- **la colinearità tra i genomi è stata mantenuta per più di 60 MILIONI DI ANNI!!!!**

Macrocolinearità tra cromosomi di mais e riso

Sono state usate 2629 ESTs di mais. Il 75% presentava sequenze ortologhe di riso

anche in regioni con elevata colinearità, attraverso processi di mappatura ad elevata risoluzione, si sono sono visti dei riarrangiamenti

quando un cromosoma di una specie è colineare con più di un cromosoma di un'altra specie, è indice di avvenute traslocazioni

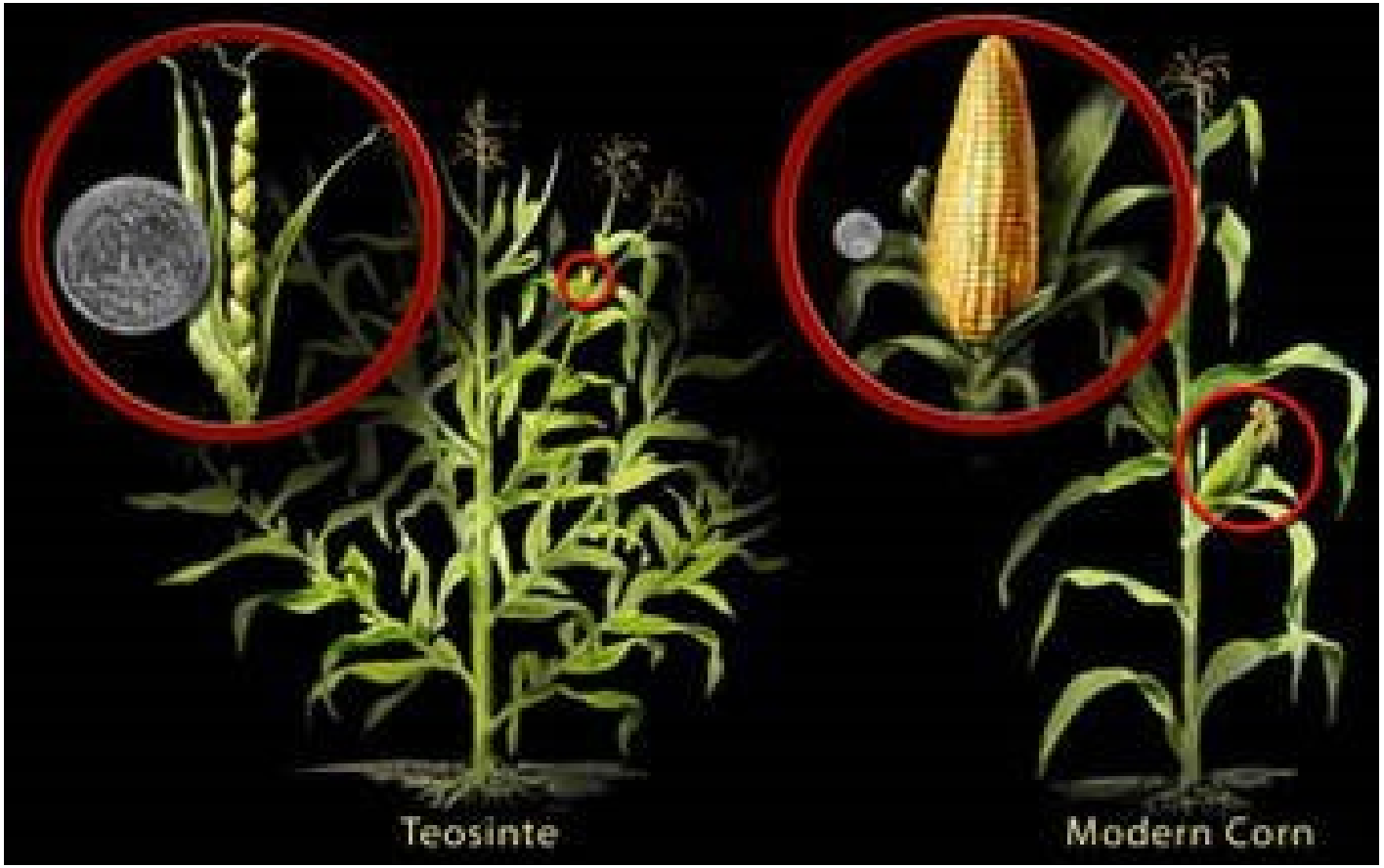


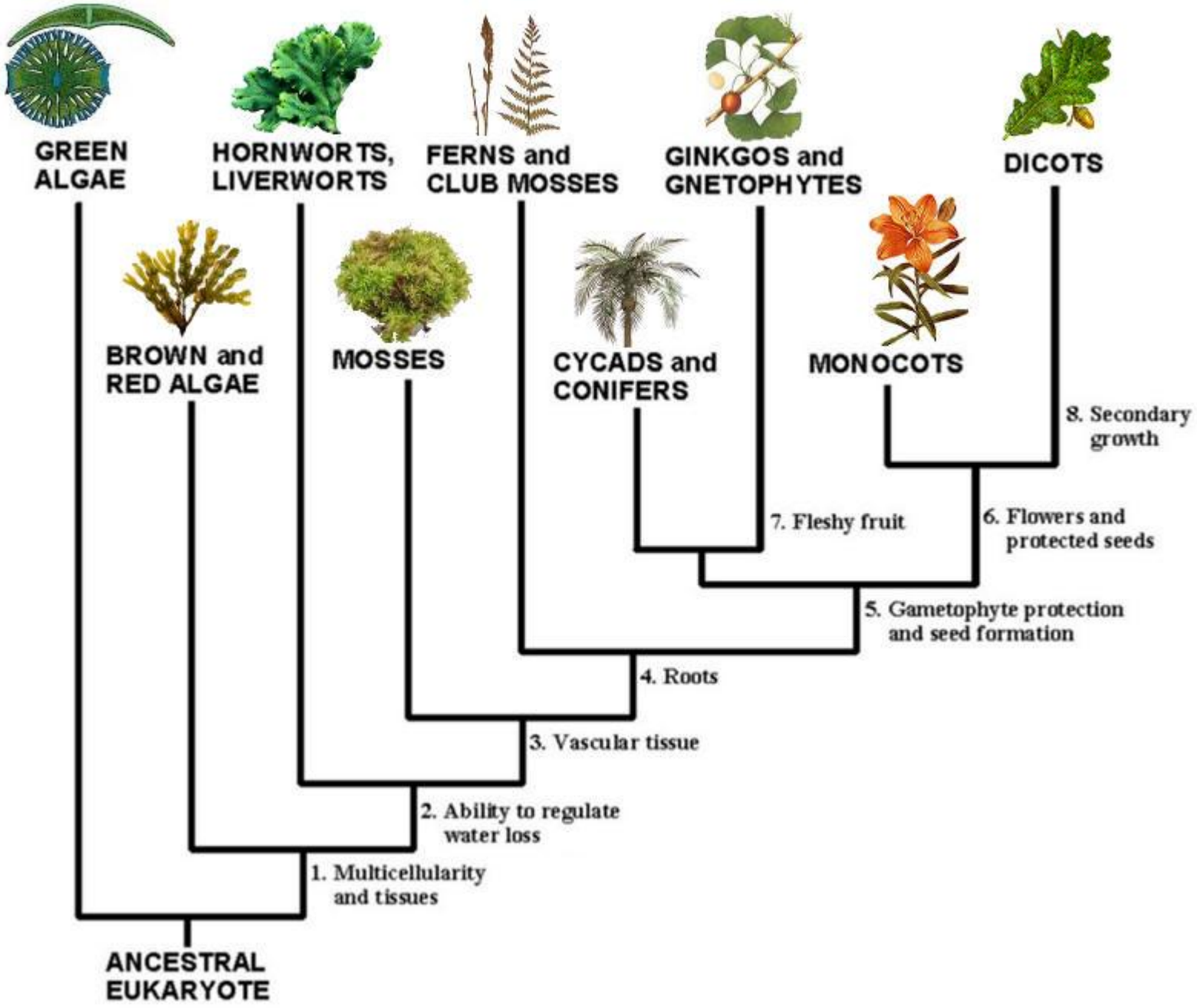
STUDIO DEI GENOMI VEGETALI

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

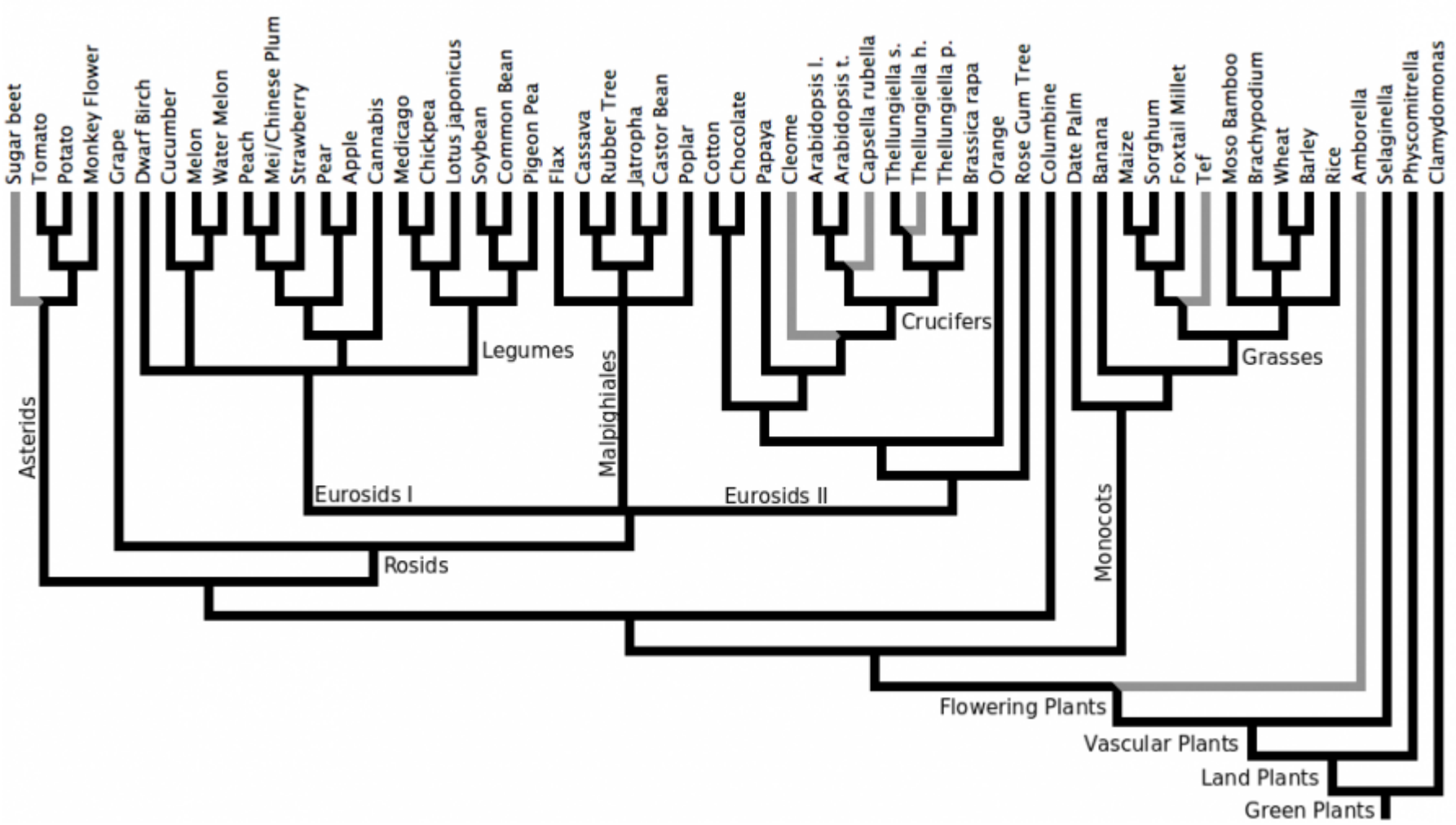
- **Comprensione dell'evoluzione delle piante**

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

Genomi di pianta sequenziati

Arabidopsis thaliana (Eukaryota) back to top

This was the first plant to be sequenced and is considered *the* species for investigating plant genetics. A member of the mustard family, the plant is popular among researchers because it grows in small spaces, lives about six weeks, and has a small genome.

» Sequenced by: The Arabidopsis Genome Initiative Abstract

» Related GNN articles:

Paranoid but Popular: Mutant mouse-ear cress offers insight into natural plant resistance

Clickable genomics: Plans for virtual plant posted

SHATTERPROOF genes in *Arabidopsis* are good news for agriculture

What makes plants grow? The *Arabidopsis* genome knows

» Image: Peggy Greb/USDA.

Oryza sativa (Eukaryota) back to top

A food staple for much of the world's population, rice comes in different varieties. Two strains were sequenced in 2002, the *japonica* (popular in Japan) and the *indica* (grown in China). An international consortium is working on a third rice genome sequence that will be the gold standard.

» Sequenced by: Syngenta and Myriad Genetics *O. sativa* L. ssp. *indica* Abstract Syngenta *O. sativa* L. ssp. *japonica* Abstract

» Related GNN article:

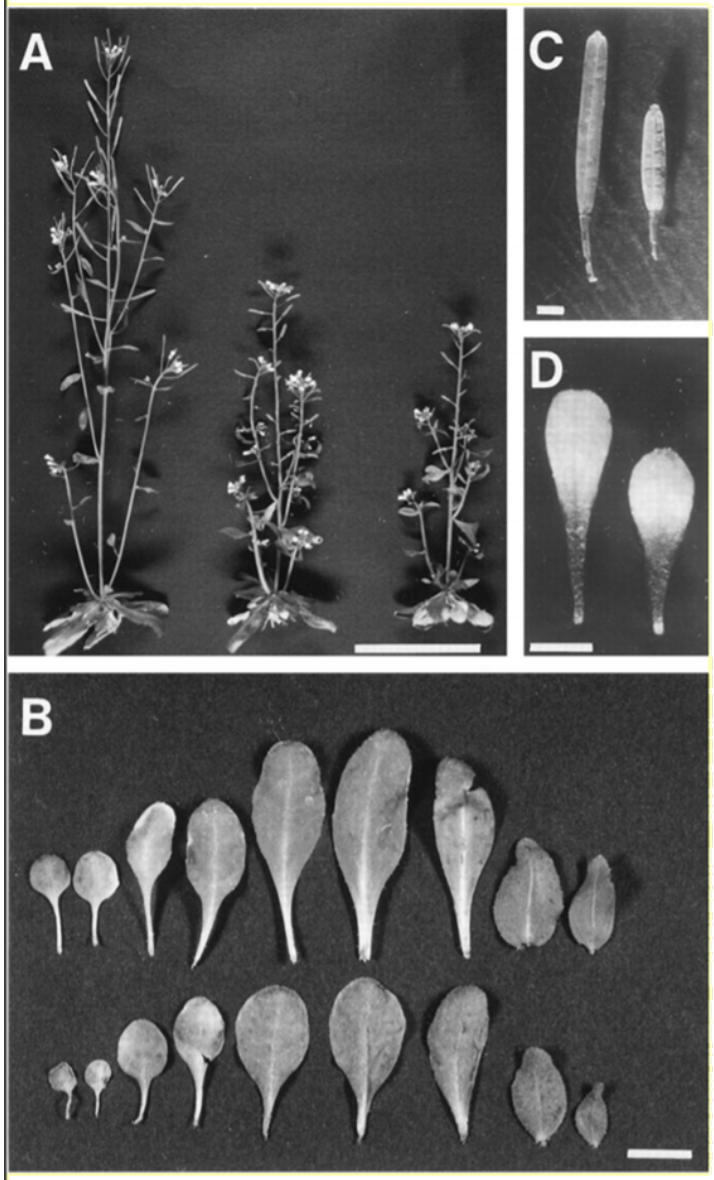
Two Groups Sequence Rice: Combining draft sequences may accelerate completion of finished genome

» Image: Photo by Ma Liwen, Courtesy of Qiu BaoXing. *(Science)*

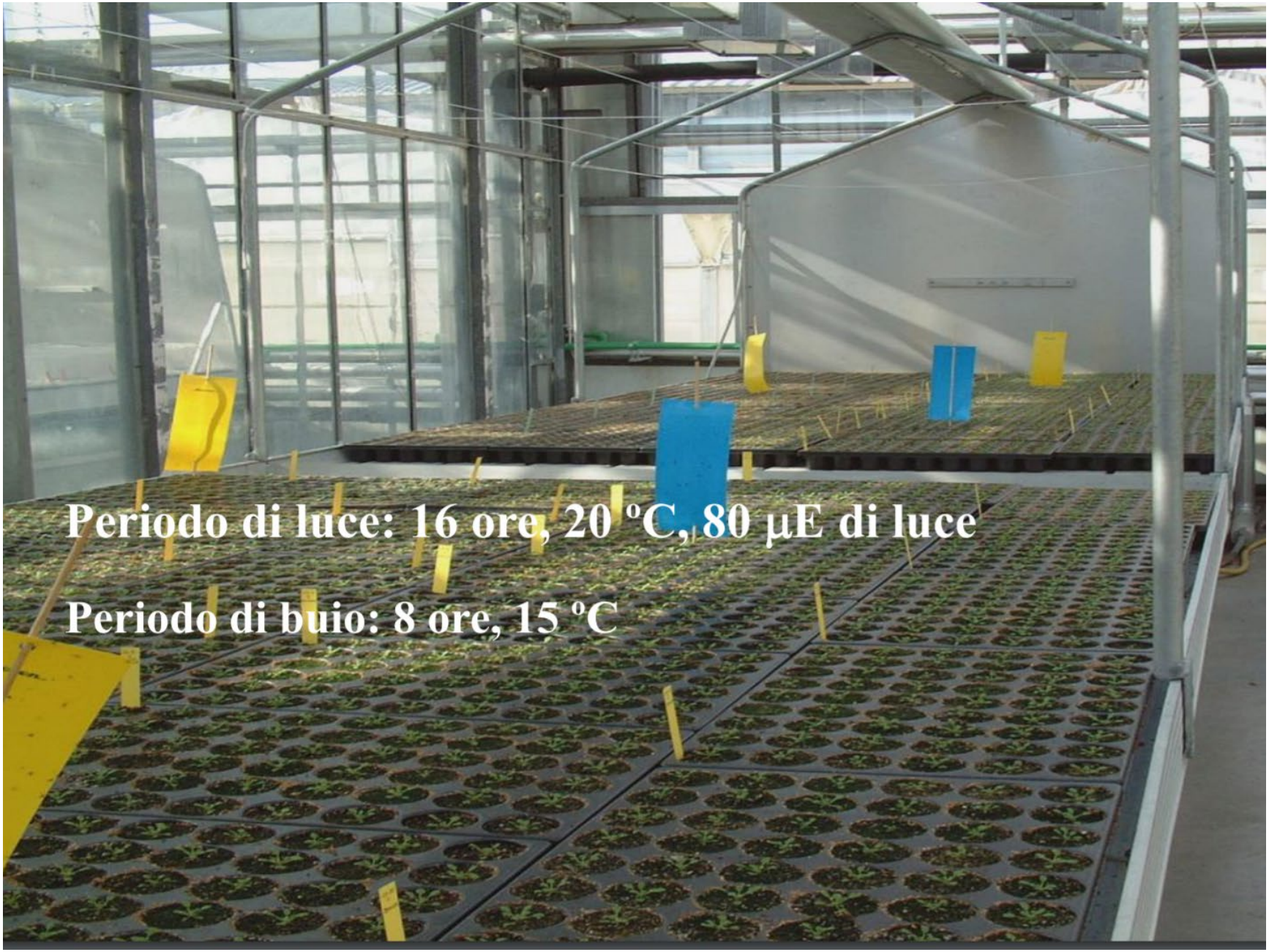
Arabidopsis thaliana

- *Arabidopsis thaliana* è una piccola angiosperma che viene usata come organismo modello nella biologia vegetale.
- *Arabidopsis* è un membro della famiglia delle Brassicaceae
- Questa pianta non ha un'importanza agronomica, ma offre importanti vantaggi per quel che riguarda la ricerca di base, in particolare per l'attribuzione di funzioni ai geni

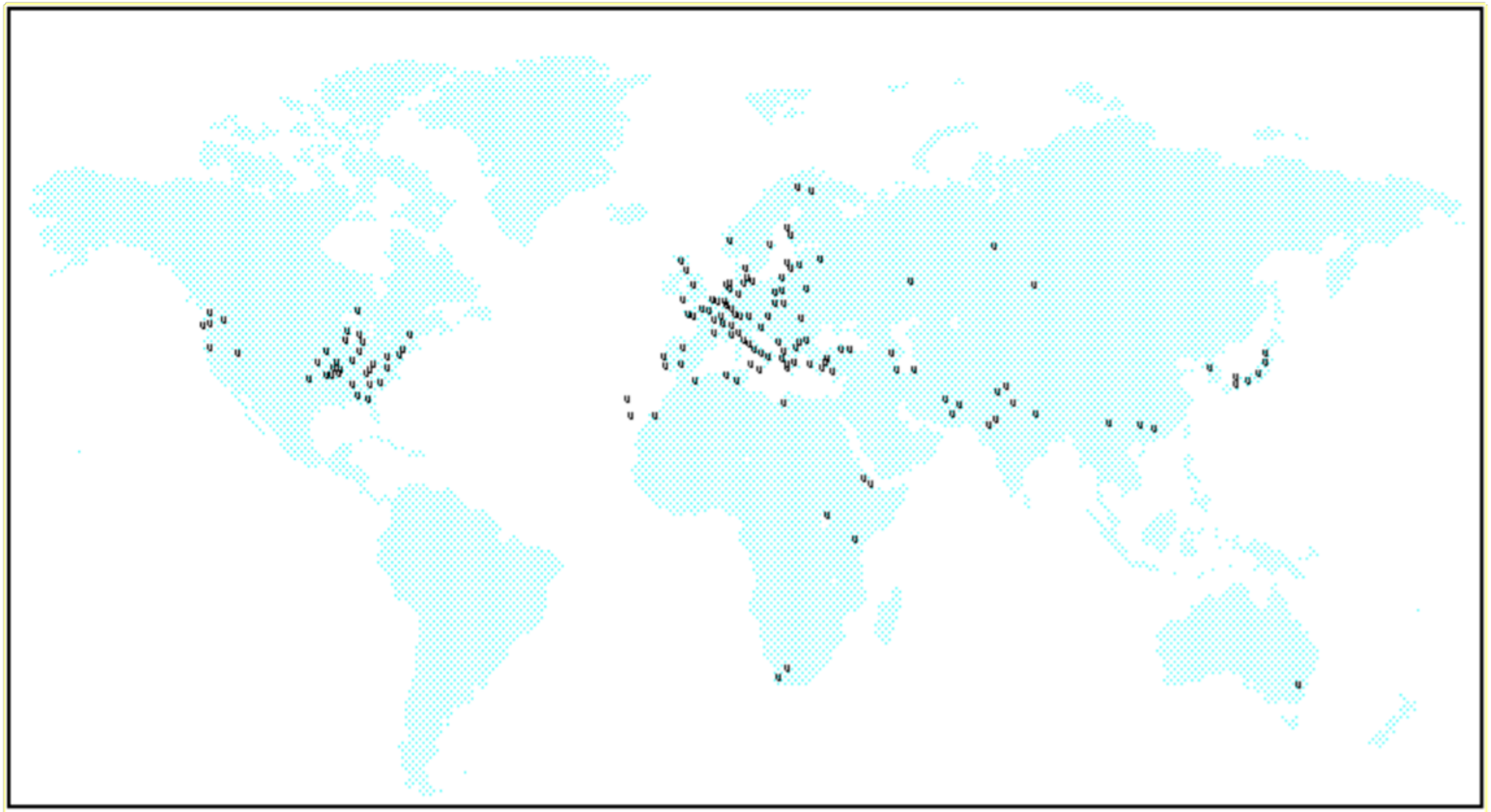
I vantaggi apportati dall'uso di Arabidopsis thaliana:



- Genoma piccolo distribuito su 5 cromosomi (114.5 Mb/125 Mb total) e diploide
- Un ciclo vitale molto rapido (circa 6 settimane dalla germinazione al seme maturo)
- E' in grado di produrre molti semi e non richiede molto spazio per essere coltivata
- Può essere trasformata utilizzando la tecnologia dell'Agrobatterio
- Disponibiltà di mutanti!!


Periodo di luce: 16 ore, 20 °C, 80 µE di luce
Periodo di buio: 8 ore, 15 °C

Distribuzione geografica delle sottospecie di Arabidopsis



Geographical distribution of ecotypes of *Arabidopsis thaliana* (L.) HEYNH.

Ampia variabilità raccolte circa 300 varietà, molto utili per la genomica funzionale, comprensione a livello molecolare di molte caratteristiche fenotipiche

- *Arabidopsis thaliana* è un'importante sistema modello per identificare geni e determinare la loro funzione. (…).

- La regione sequenziata comprende 115.4 Mbp delle 125 Mbp che costituiscono il suo genoma. (…).

- Il genoma contiene 25,498 geni che codificano proteine appartenenti a 11.000 famiglie, mostrando una diversità funzionale comparabile a quella di *Drosophila* e *C.elegans* (…).

- è il primo genoma sequenziato di una pianta e fornisce le basi per comparare processi conservati in tutti gli organismi eucarioti.

- Permetterà di identificare le funzioni di geni specifici per le piante e individuare geni importanti per il miglioramento delle piante di valore commerciale

MISSION STATEMENT

To exploit the revolution in plant genomics by understanding the function of all genes of a reference species within their cellular, organismal, and evolutionary context.

Il progetto 2010 ha lo scopo di comprendere la funzione di tutti i 25,000 geni identificati nel genoma di Arabidopsis…

Il fine ultimo del progetto è conoscere ogni aspetto molecolare dello sviluppo di una pianta, al punto tale da poter simulare la crescità di una pianta virtuale…

…Alla fine saremo in grado di predire la funzione di geni appartenenti a specie di interesse agronomico, attraverso il confronto delle loro sequenze

Quando la funzione di tutti i geni sarà nota saremo in grado di:

1. Predictable outcomes to directed experimental genetic changes.
2. Directed genetic changes that accelerate domestication of wild species.
3. Facile genetic manipulation that ensures maintenance of, and expansion of germplasm bases.
4. A description of the underlying mechanisms of heterosis, and the ability to use this phenomenon more effectively.
5. Enhanced understanding of the genetic basis of phenotypic plasticity, which will have a profound impact not just in plants, but also in animals, including humans.
6. Knowledge of the minimum gene set required for plant life.
7. Understanding of the genetic basis of plant evolution, which will enrich our understanding of the diversity of life on earth.
8. An understanding of interactions between plants and other organisms in their environment, up to the level of ecosystems.

1. **predire gli effetti di una qualsiasi modificazione genetica**
2. **modificare a nostro vantaggio specie selvatiche**
3. **Mantenere ed espandere l'insieme dei**
4. **germoplasmi**
4. **comprendere i meccanismi alla base del fenomeno indicato come Eterosi**
5. **comprendere le basi molecolari alla base della plasticità dei fenotipi**
6. **conoscere il minor numero di geni necessari alla vita della pianta**
7. **comprendere le basi genetiche dell'evoluzione**
8. **comprendere le basi molecolari delle interazioni tra le piante ed altri organismi**

IDENTIFICARE LA FUNZIONE DI UN GENE

La vastità del numero di sequenze oggi disponibili ha spostato l'interesse della ricerca verso l'identificazione delle funzioni geniche.

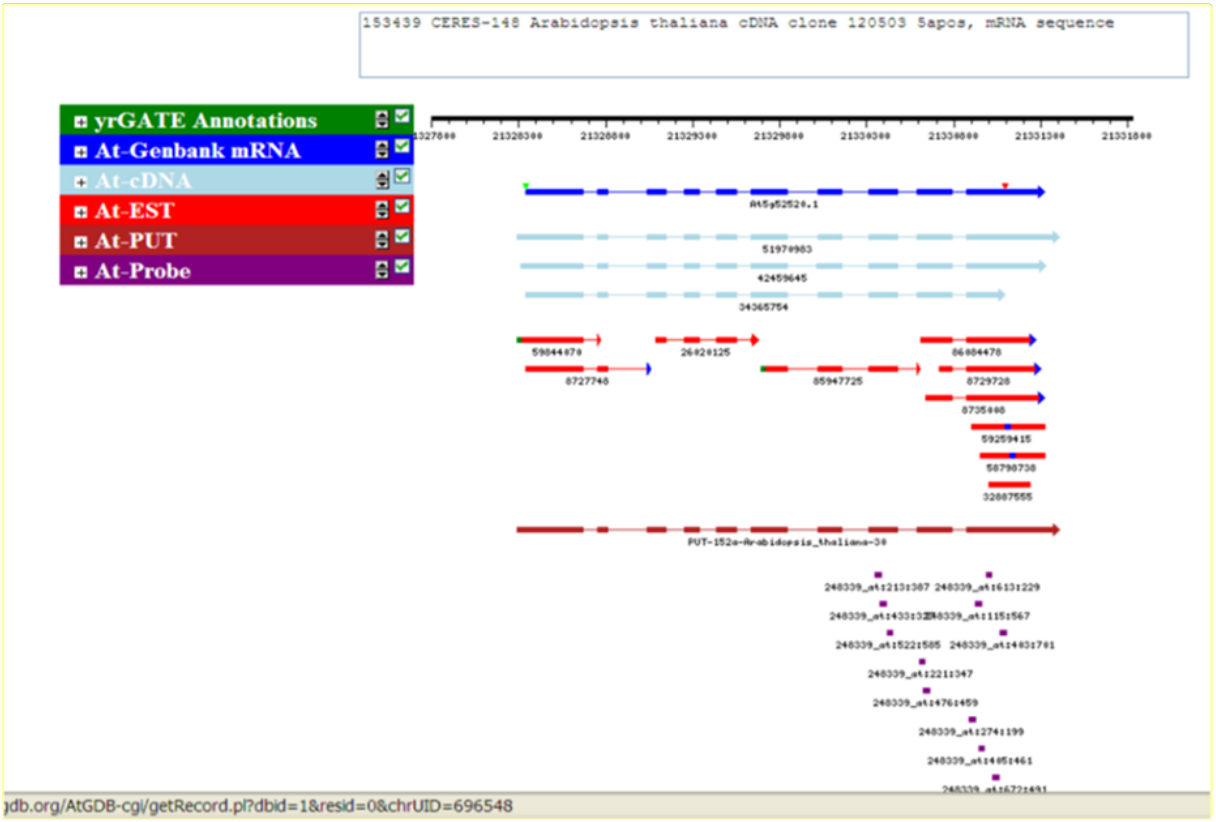
In un genoma completamente sequenziato non si sa, dalla sequenza semplice, quali siano i geni e quali siano le loro funzioni.

Lo scopo finale della genomica funziona è la definizione del rapporto esistente tra gene e fenotipo:

non è sufficiente definire la funzione biochimica di ciascuna sequenza proteica ma è comunque necessario definire le interazioni tra i diversi prodotti genici

Come si individua un gene e la proteina da esso codificata all'interno di una sequenza genomica

- Si utilizza la sequenza genomica per individuare le molecole di cDNA o anche ESTs (Expressed sequence tags) corrispondenti



Come si individua un gene e la proteina da esso codificata all'interno di una sequenza genomica

- Si utillizza la sequenza genomica per individuare le molecole di cDNA corrispondenti

- Si utilizza la sequenza della molecola di cDNA per suddividere la sequenza genomica in esoni (sequenze codificanti) ed introni (sequenze non codificanti)

Come si individua un gene e la proteina da esso codificata all'interno di una sequenza genomica

- Si utillizza la sequenza genomica per individuare le molecole di cDNA corrispondenti

- Si utilizza la sequenza della molecola di cDNA per suddividere la sequenza genomica in esoni (sequenze codificanti) ed introni (sequenze non codificanti)

- All'interno della sequenza di cDNA si individuano il codone di inizio della traduzione (AUG) e il codone di stop (UAA, UAG, UGA) cioè la CDS

- La traduzione della sequenza CDS risulterà in una sequenza a.a. Il confronto della sequenza aminoacidica con altre sequenze proteiche può permettere l'individuazione della funzione della proteina in esame

Arabidopsis thaliana

Dicotiledone (*Brassicaceae*)

Piccolo genoma dipolide (C1 = 125 Mbp)

Trasformabile facilmente

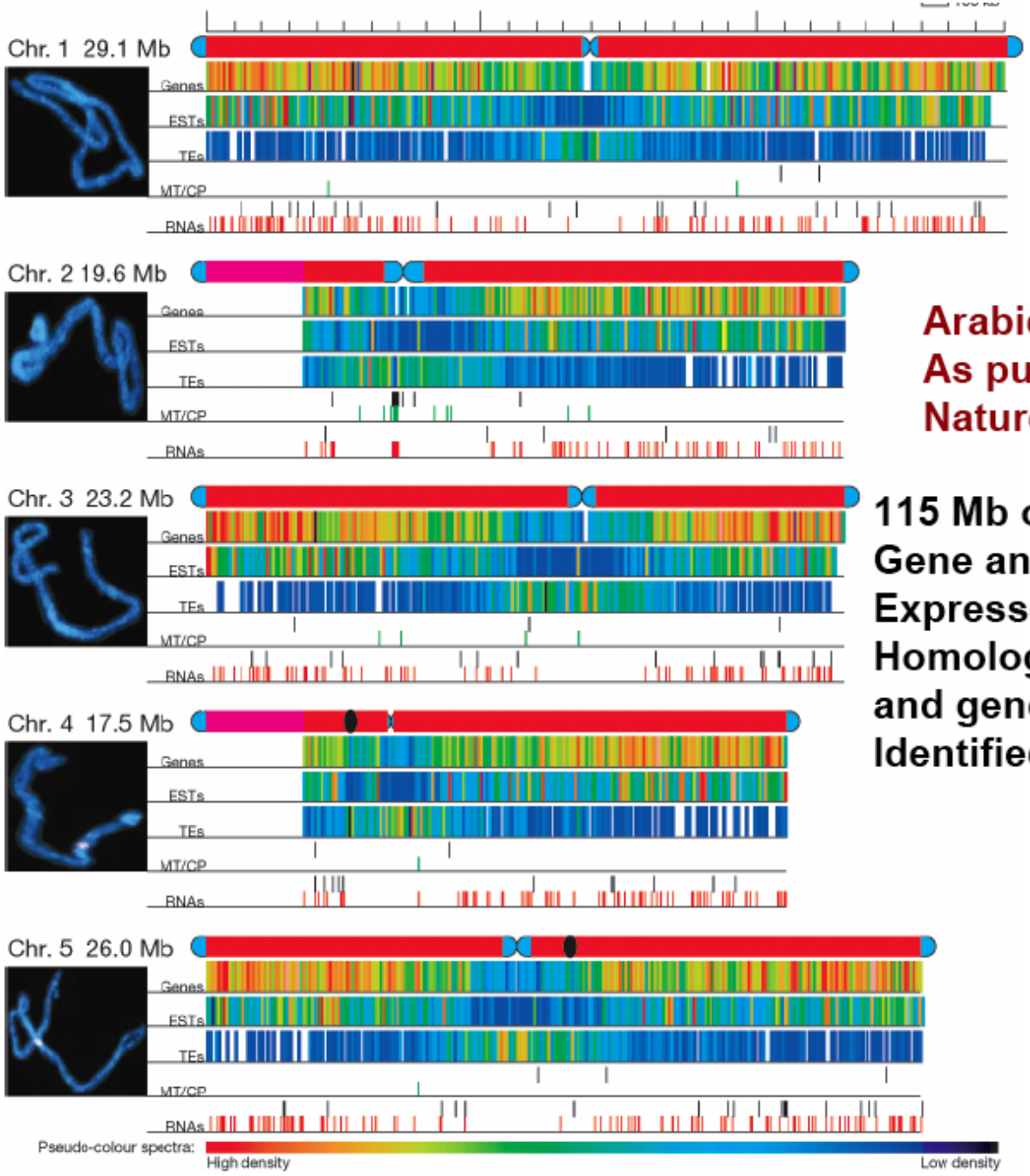
5 cromosomi

Piccole dimensioni

Ciclo vitale breve (2 mesi)

Primo genoma sequenziato

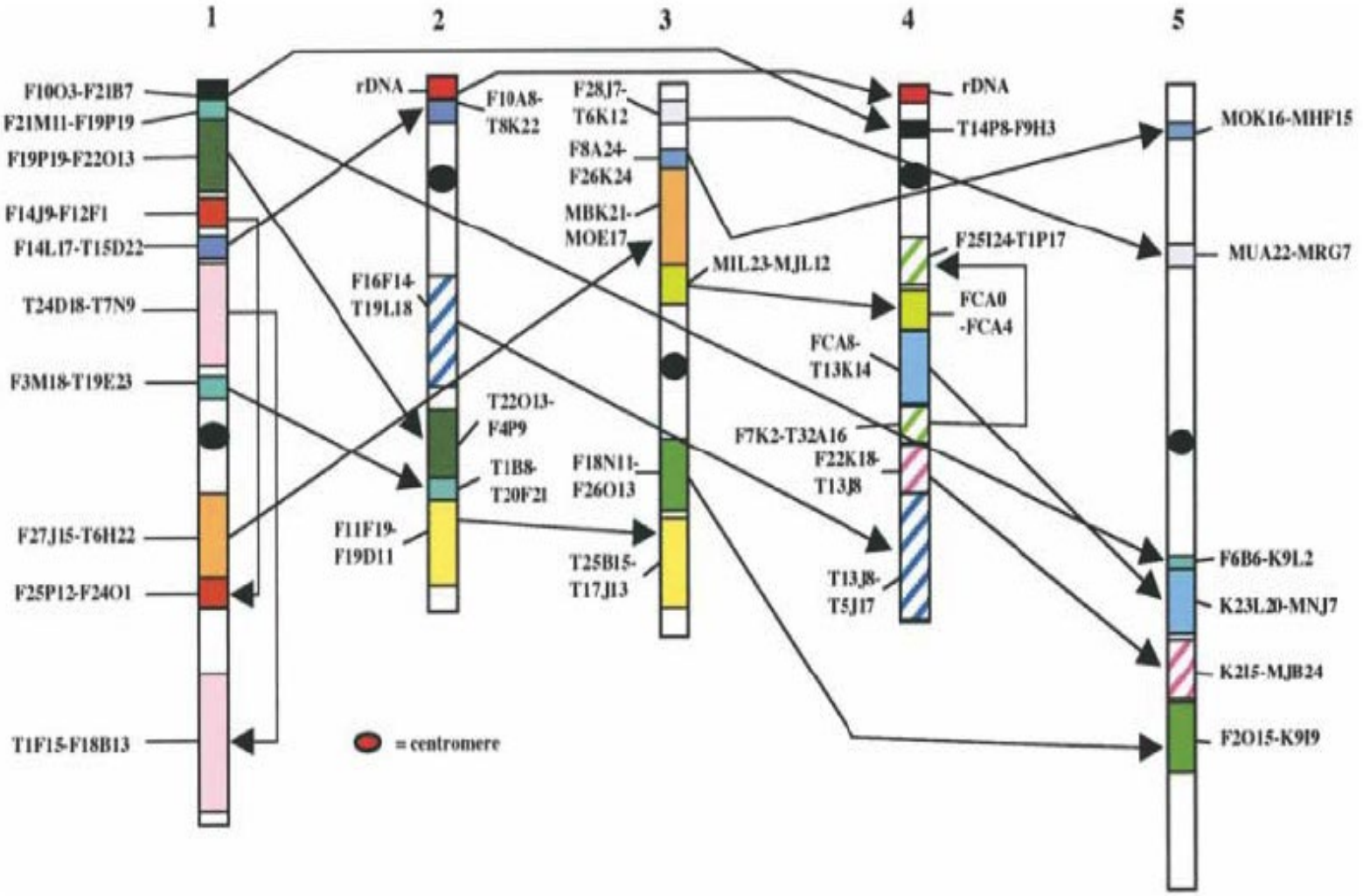
The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408 (6814), 796-815 DOI: 10.1038/35048692



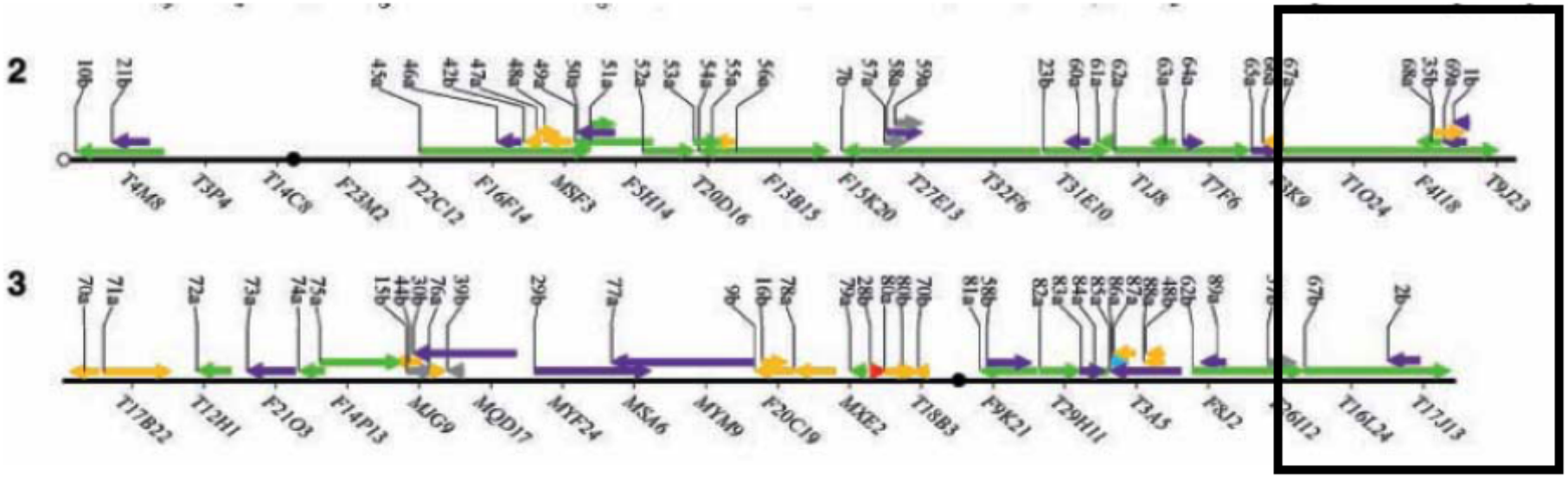
Arabidopsis genome sequence.
As published in 2000.
Nature 408, 796.

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.

SHATTERPROOF 1

100% identical in MADS DNA binding domain

SHATTERPROOF2

In Arabidopsis

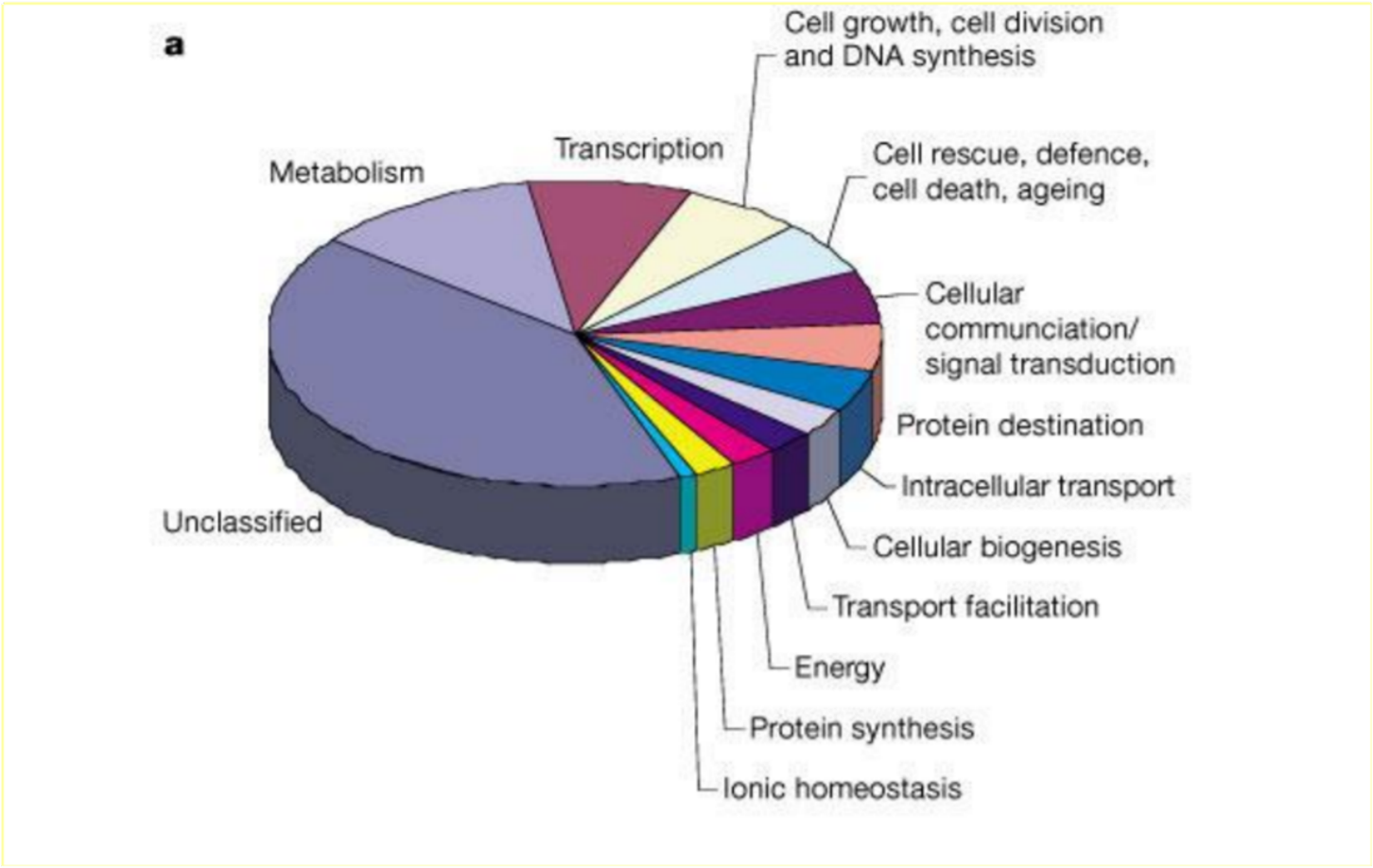
La funzione del 69% dei geni è stata identificata per 'omologia' o 'similarità' con proteine a funzione nota

Solo il 9% è stato caratterizzato sperimentalmente

Molti fattori di trascrizione in Arabidopsis hanno avuto un'evoluzione indipendente da quella di altri eucarioti

Il genoma presenta più omologia con quello di organismi pluricellulari che non con il lievito (trasduzione del segnale, comunicazione cellulare)

Classificazione dei geni di Arabidopsis in categorie funzionali



Quasi la metà dei geni di Arabidopsis sono stati classificati come codificanti per proteine sconosciute.

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%)

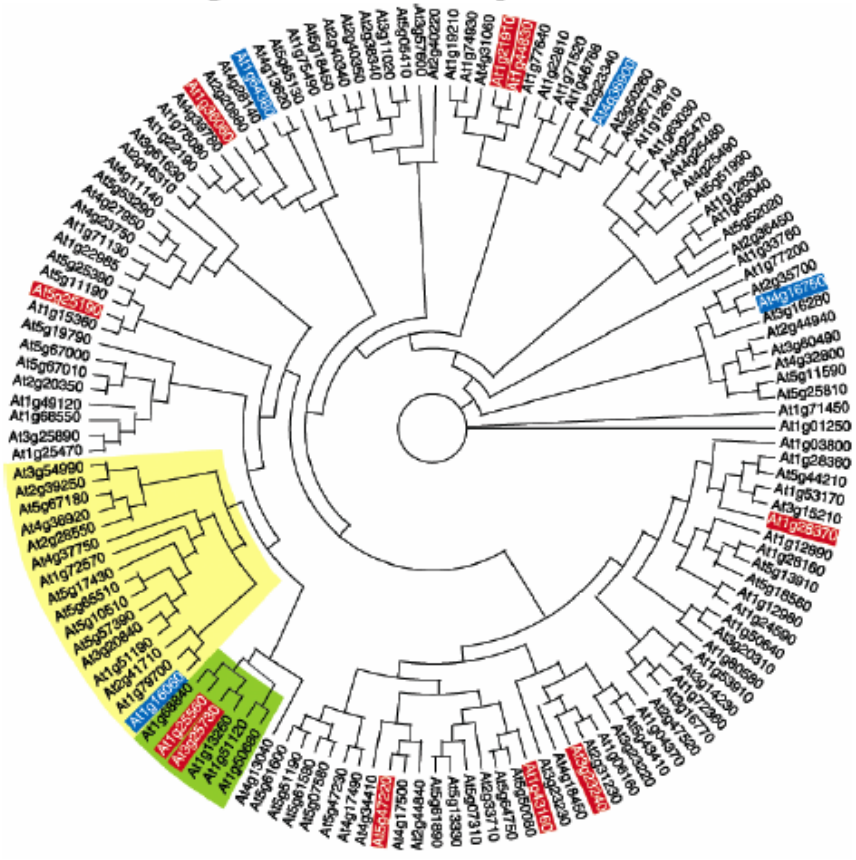
Drosophila contains around 640 genes encoding transcription factors, around 4.5%.

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.

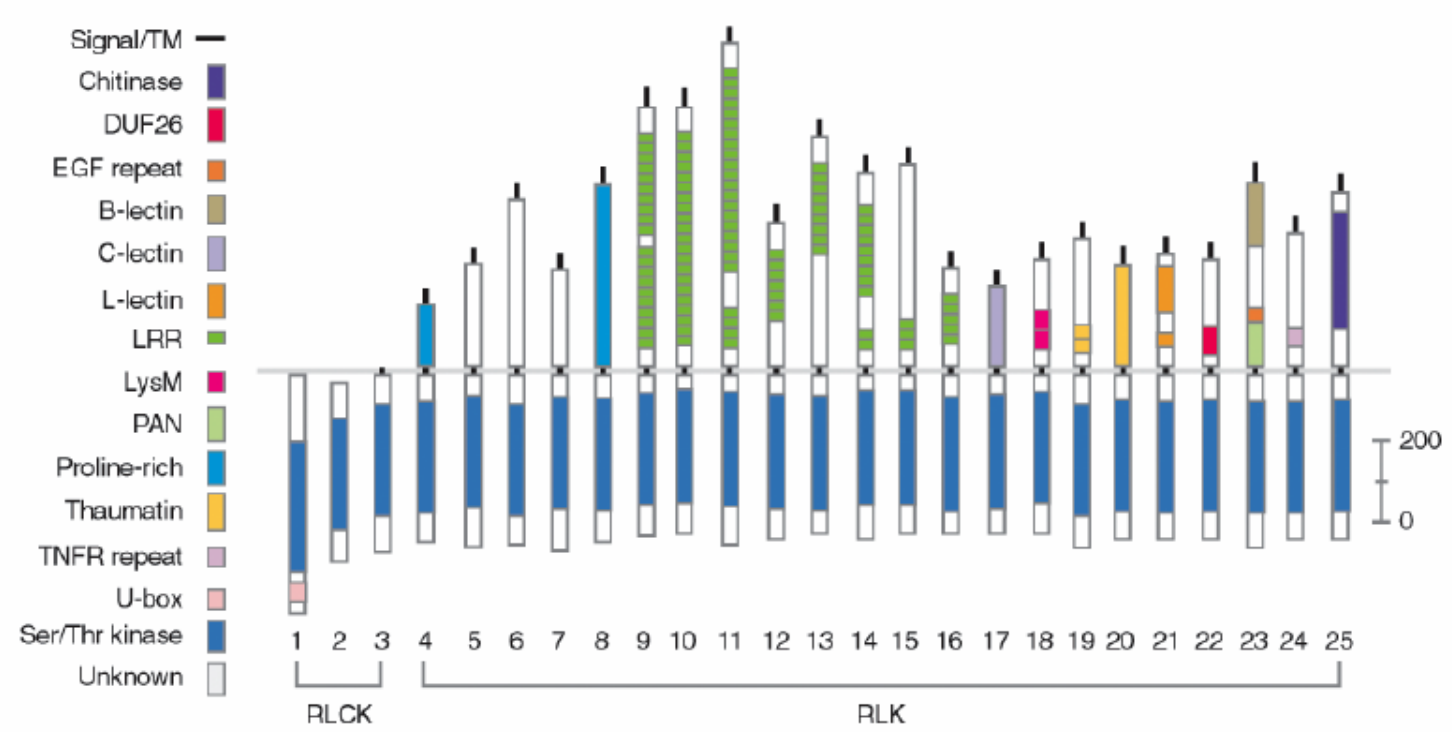


141 AP2-domain TFs in Arabidopsis

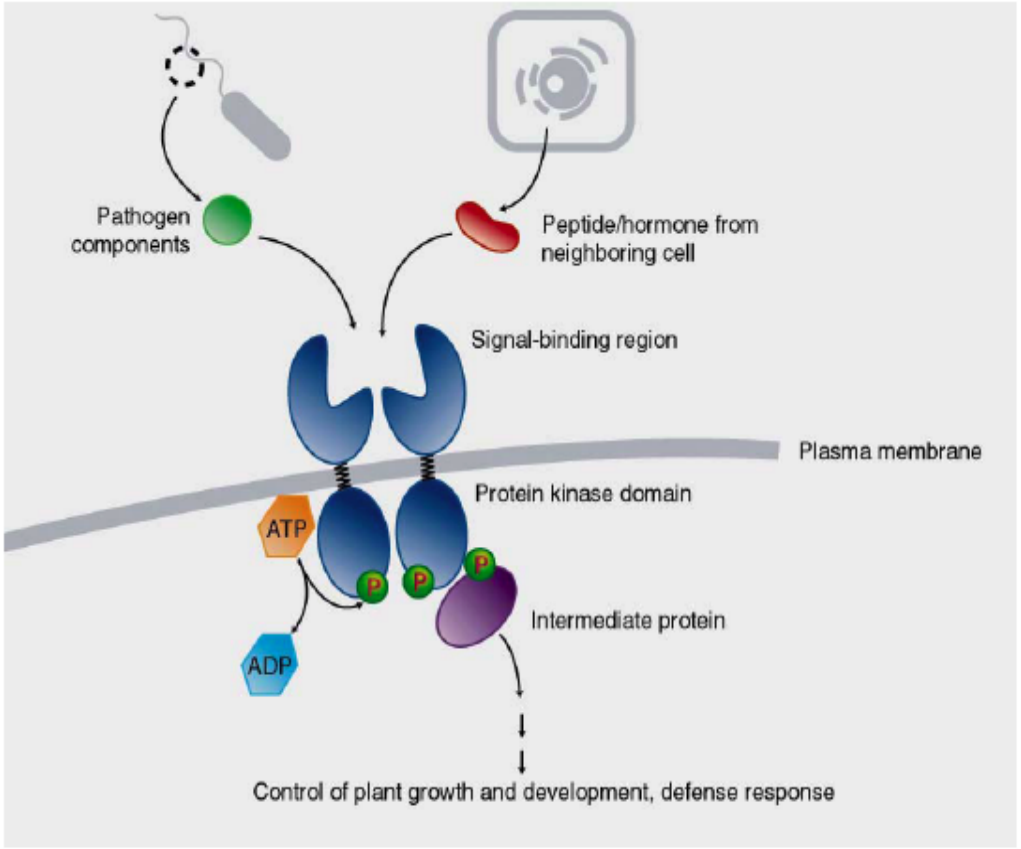
-Some related to APETALA2, Involved in flower development (yellow)
-Some involved in ethylene response and their transcription is changed by ethylene (red/blue)
-Others are involved in stress responses to drought

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

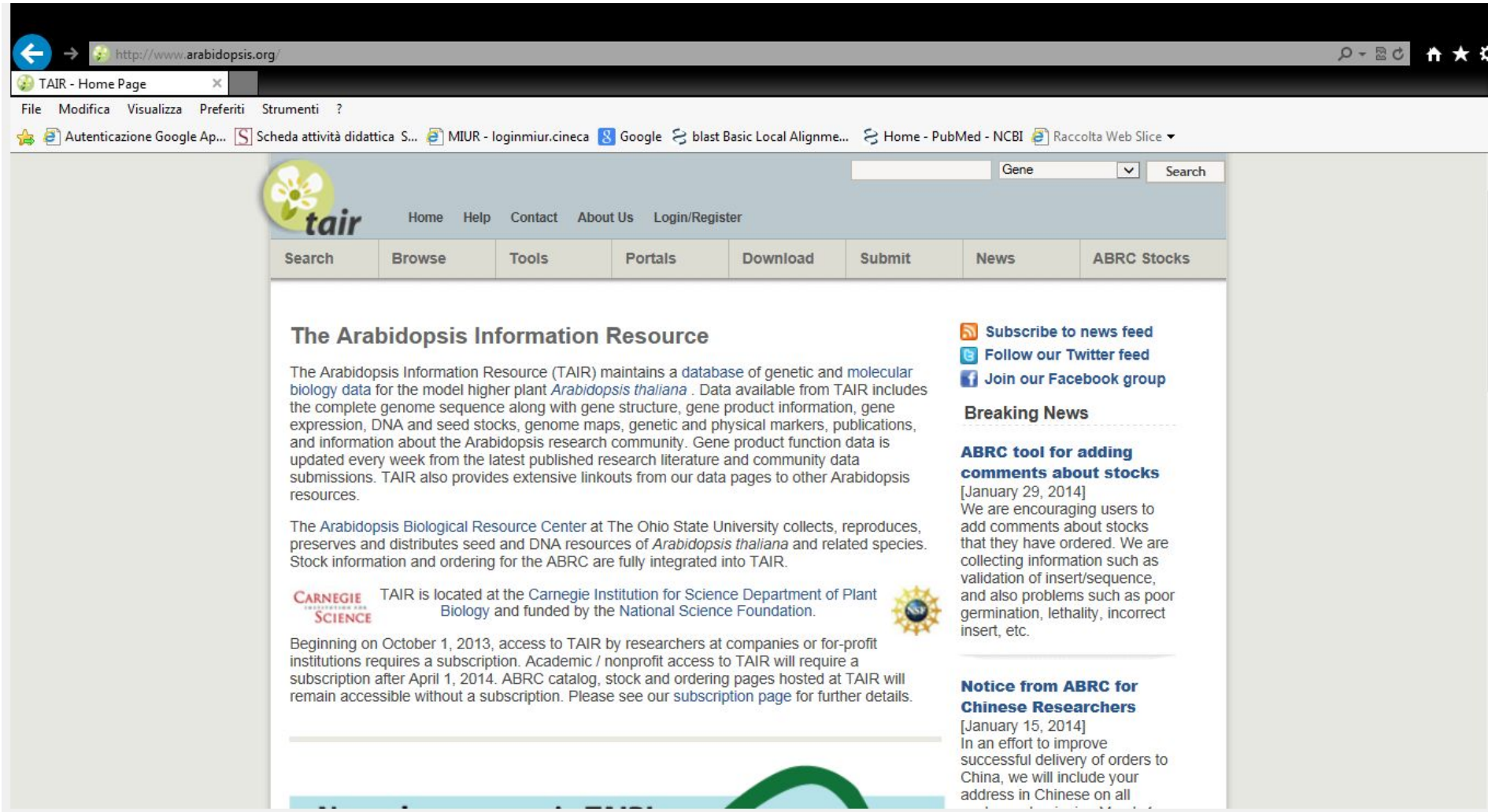
- **light harvesting**
- **chlorophyll biosynthesis**
- **carbon dioxide fixation**
- **energy generating photosystems**

Transporters

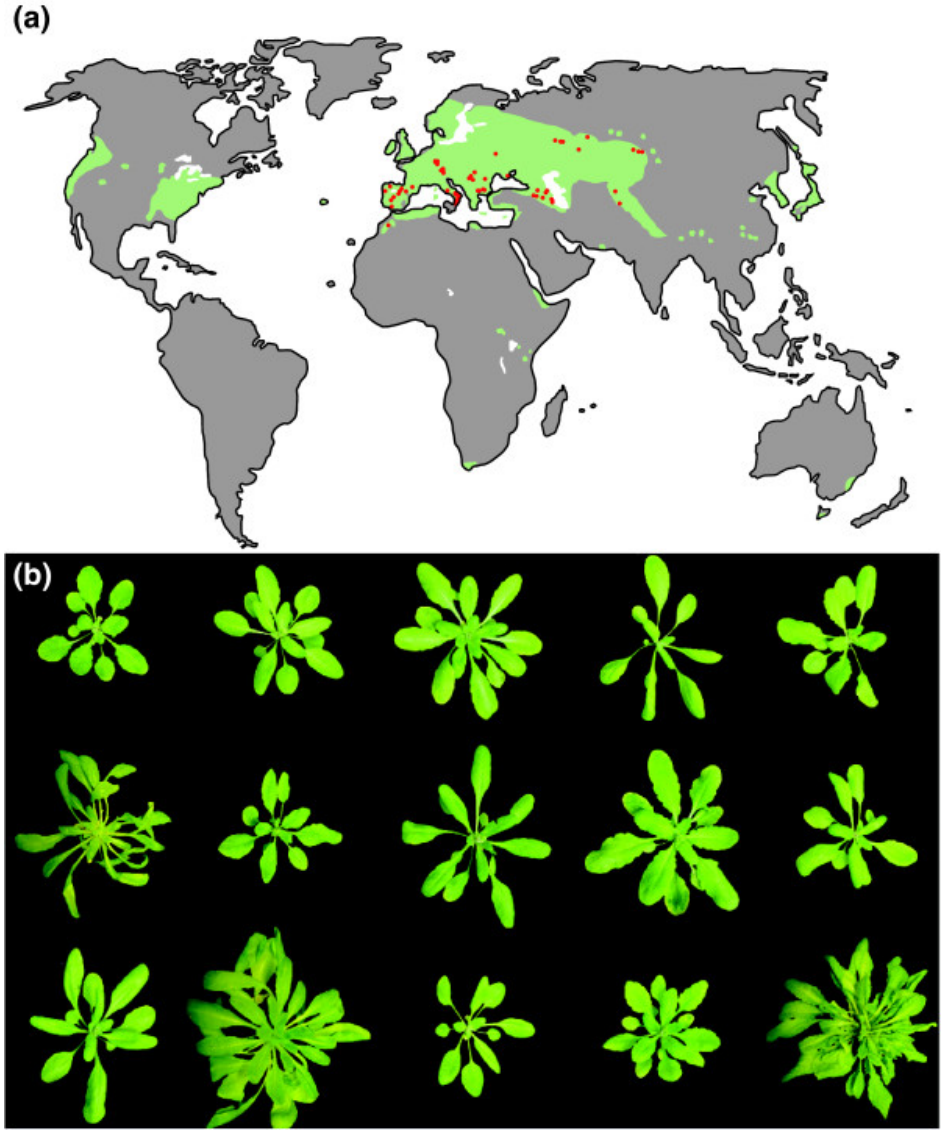
- **plants mainly use proton-type ATPases, whereas in animals transport is usely coupled with sodium ions via sodium-type ATPases.**

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

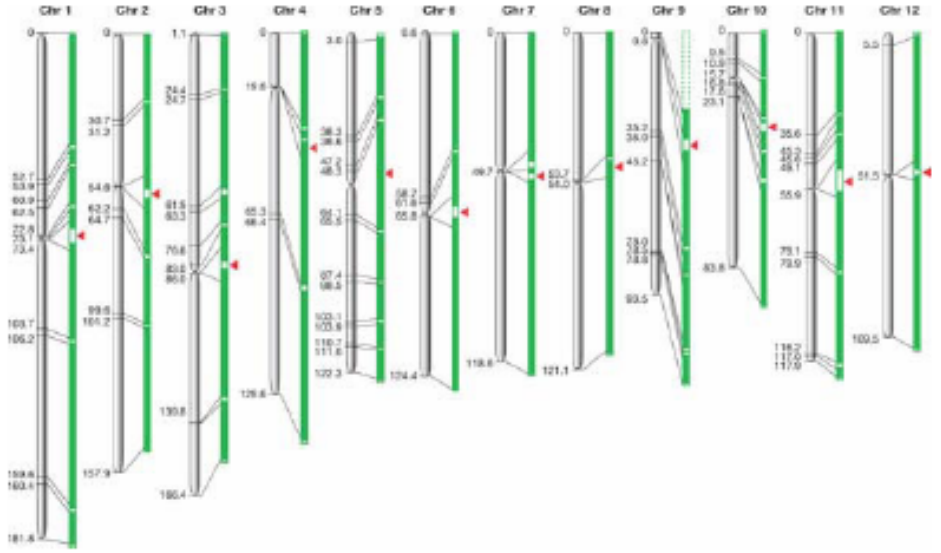


818 accessioni sequenziate e rilasciate al 13-3-2014

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)

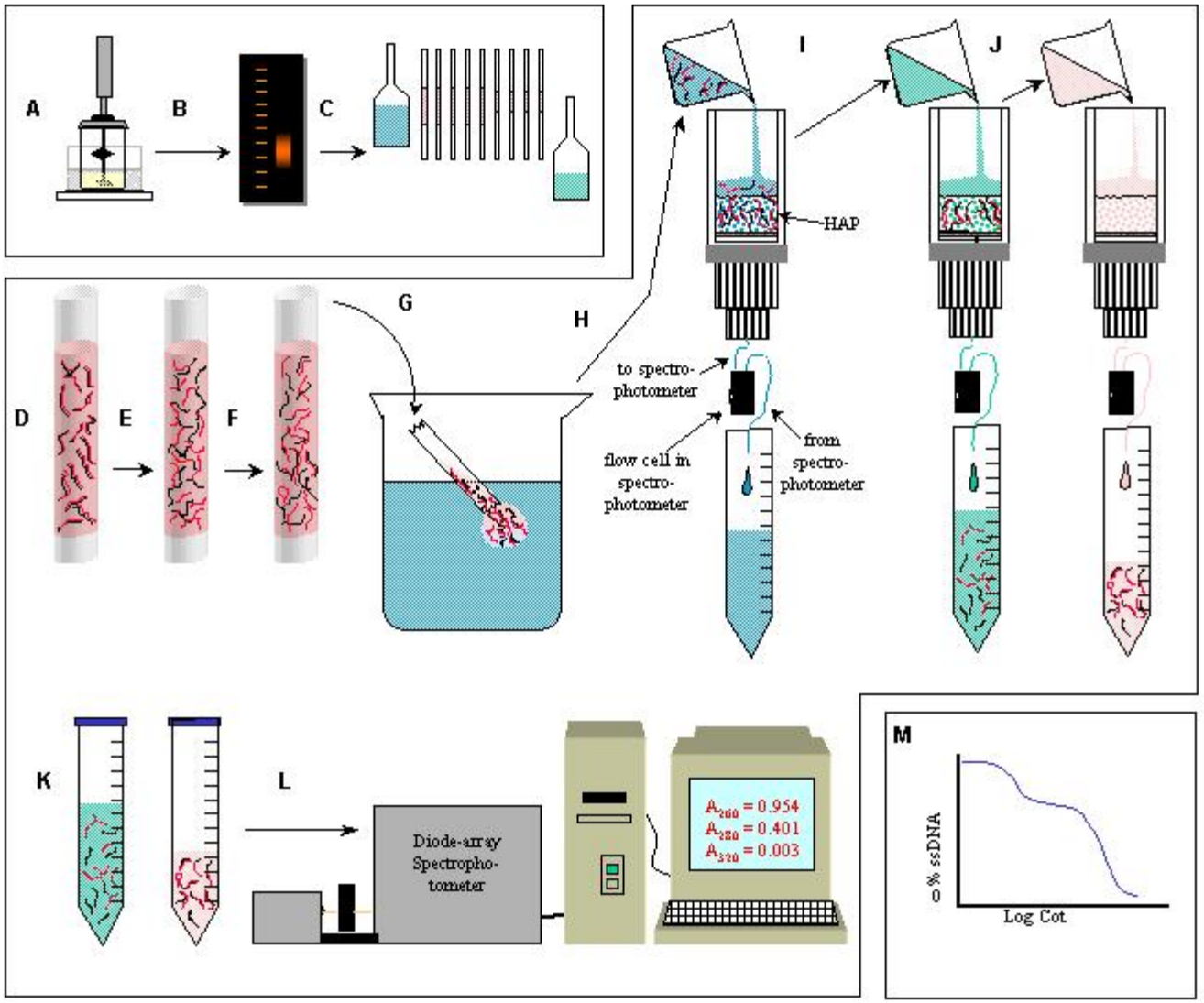
DNA ripetitivo

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

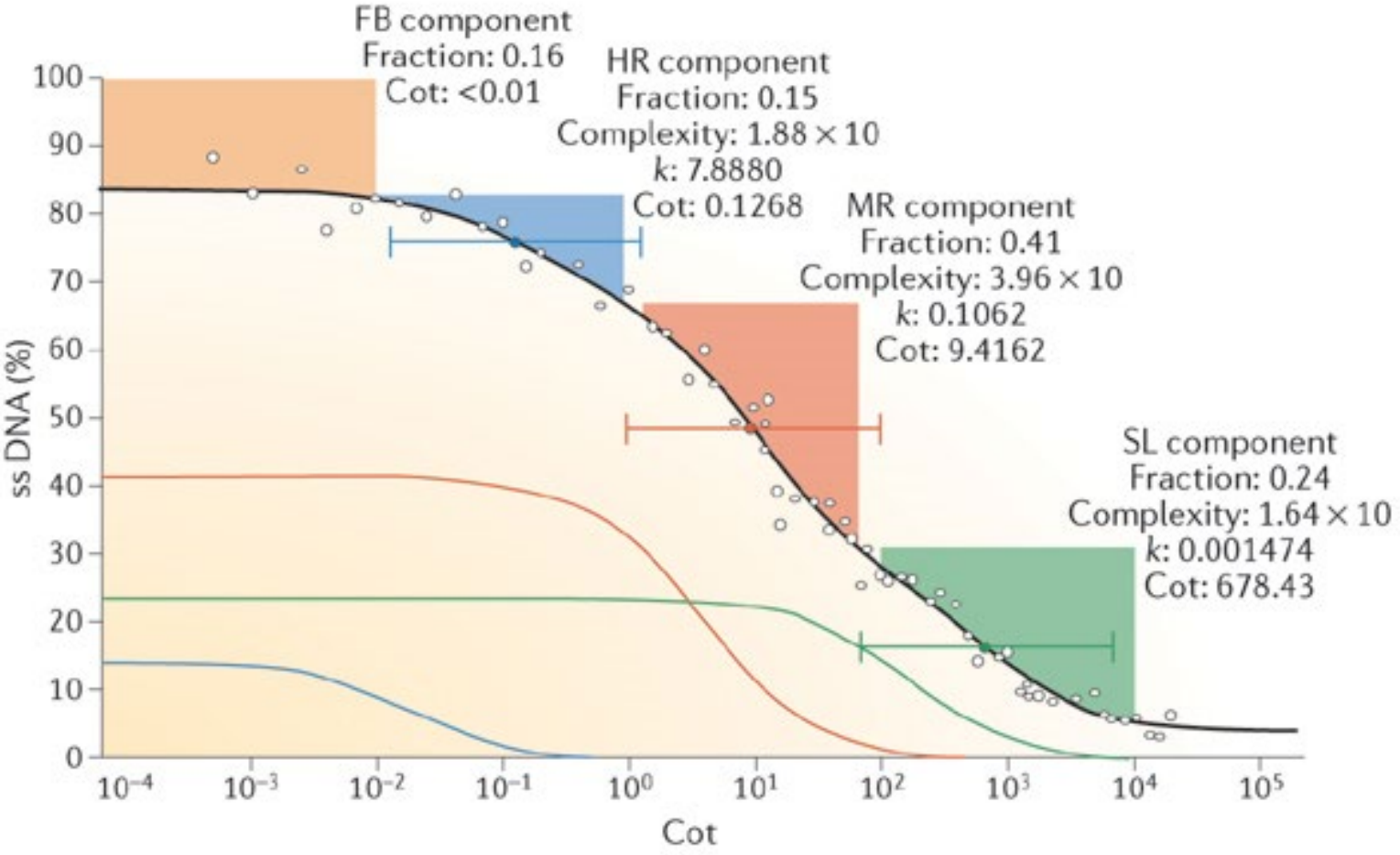
Cinetica di riassociazione

- 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



Cinetica di riassociazione



Nature Reviews | Genetics

Clonaggio basato sul valore Cot (CBCS)

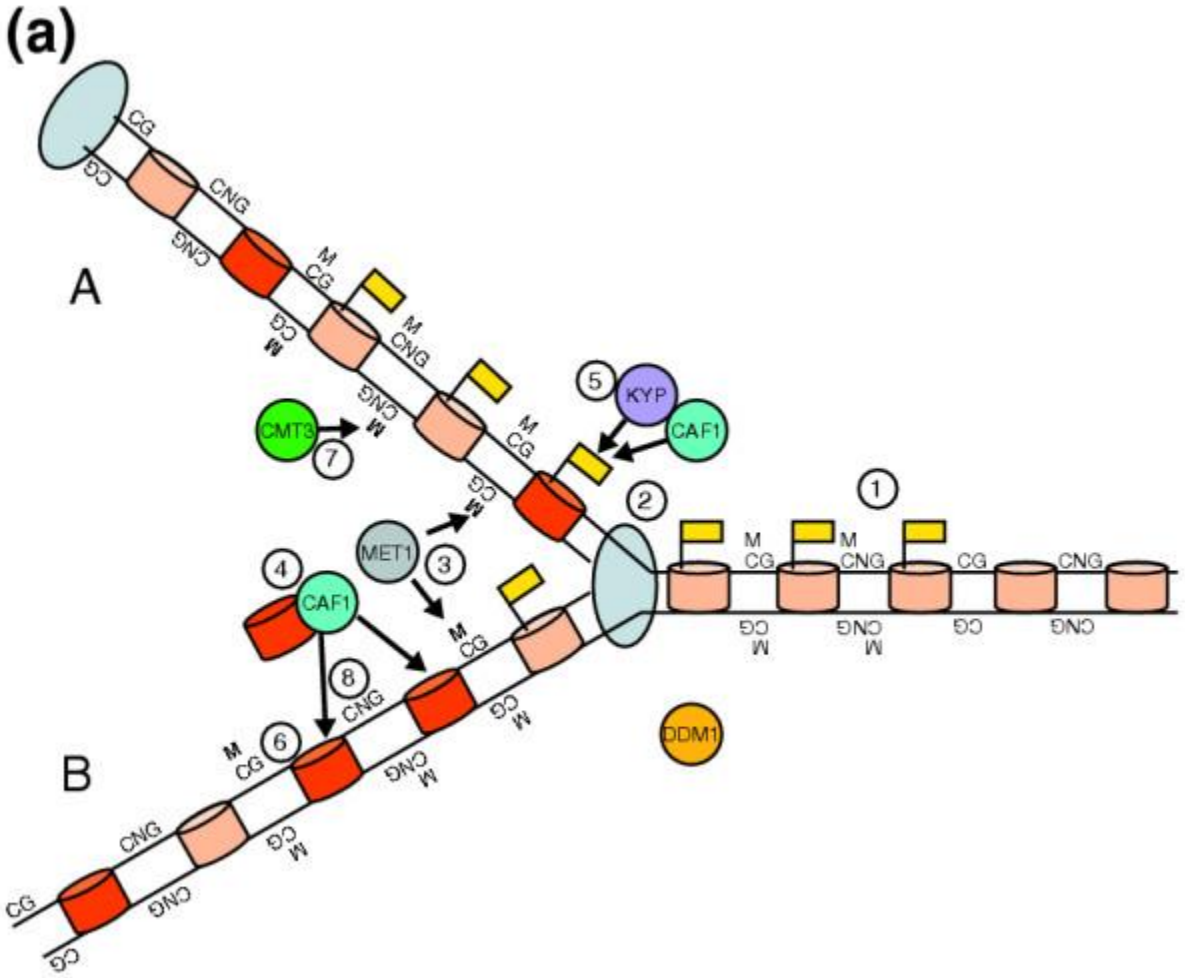
L’analisi Cot permette di isolare specificamente frazioni più o meno ripetitive

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)

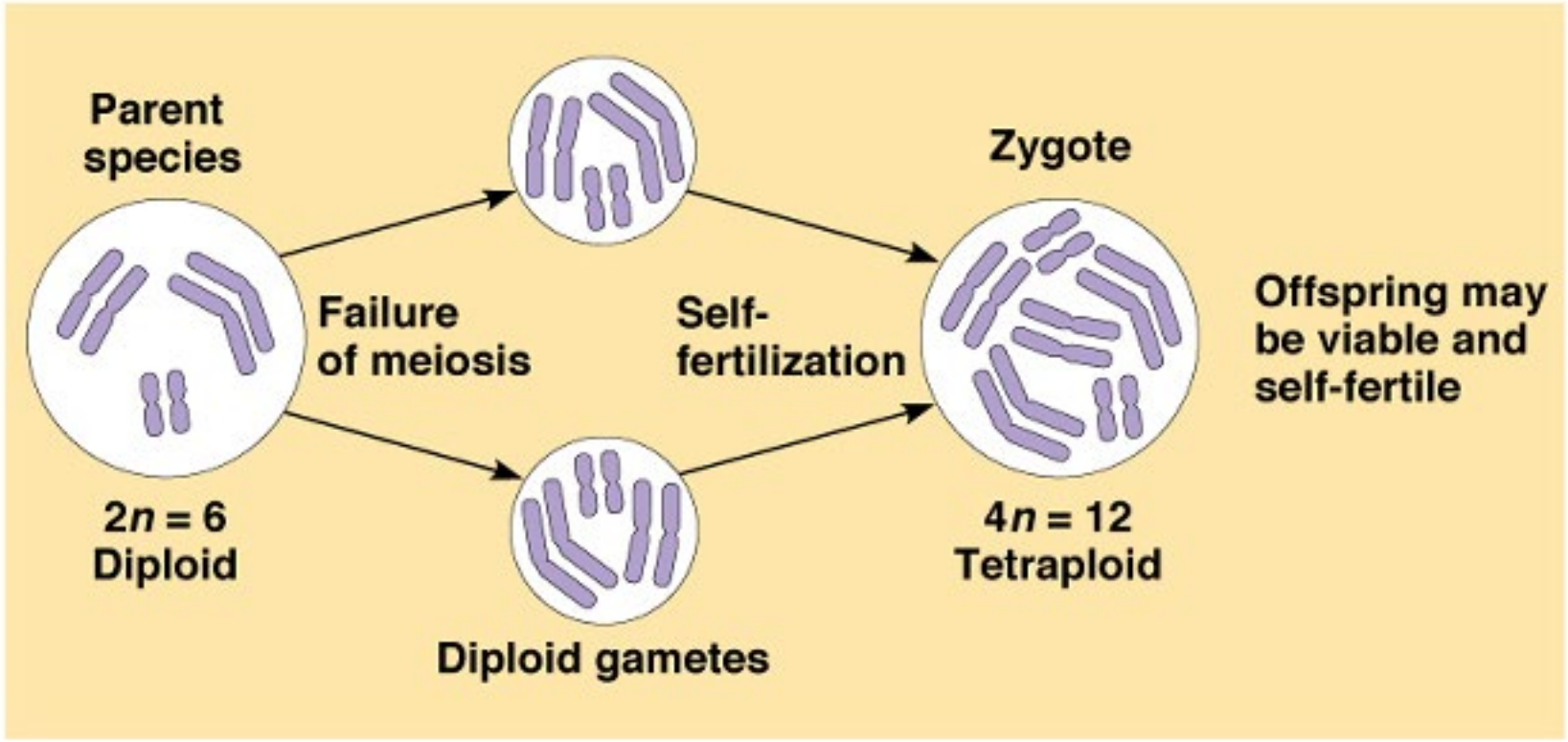
POLIPLOIDIA

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

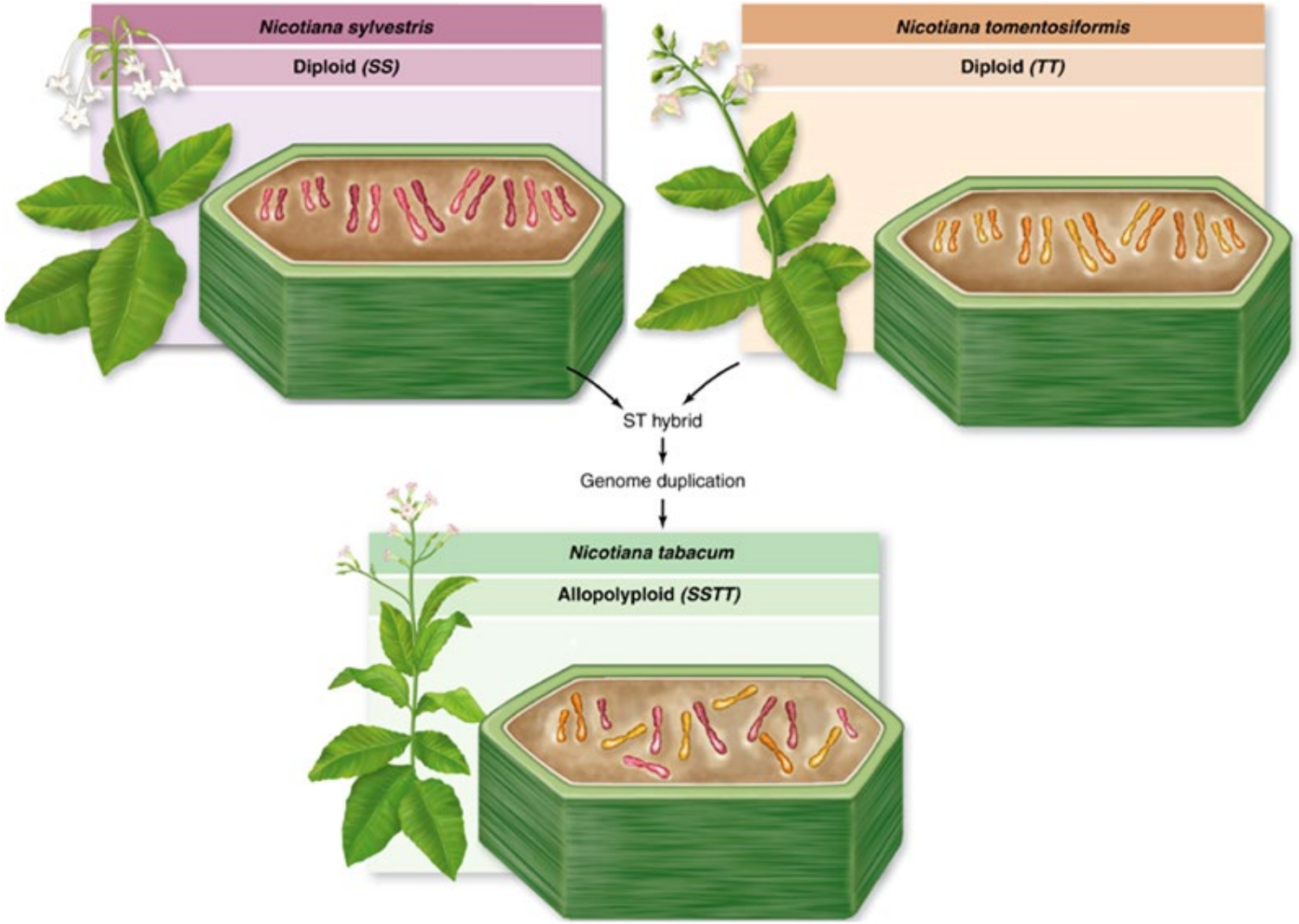
- Ibridazione di due specie diverse (**allopoliploidia**)

AUTOPOLIPLOIDIA

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



ALLOPOLIPLOIDIA: TABACCO



ALLOPOLIPLOIDIA: FRUMENTO

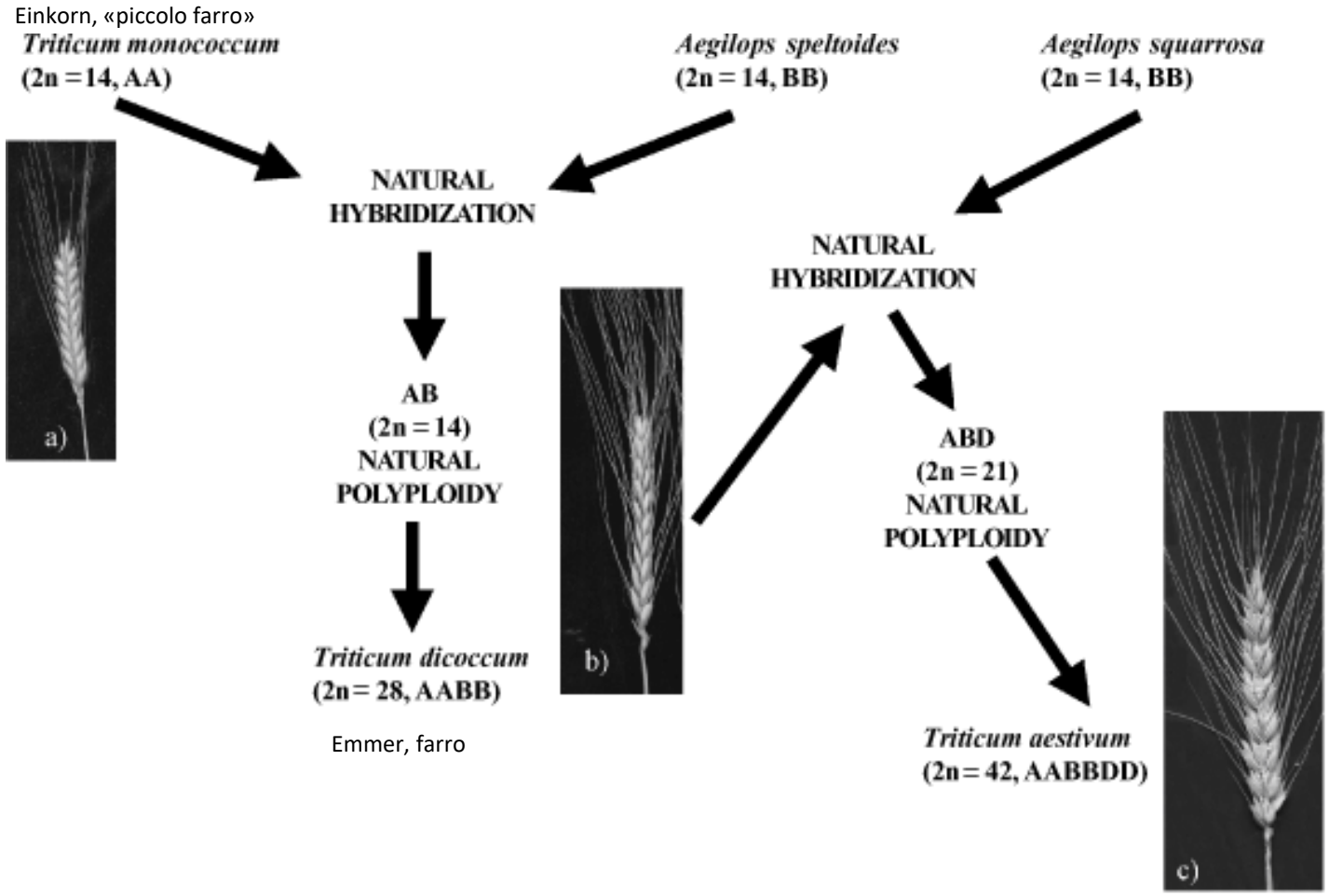
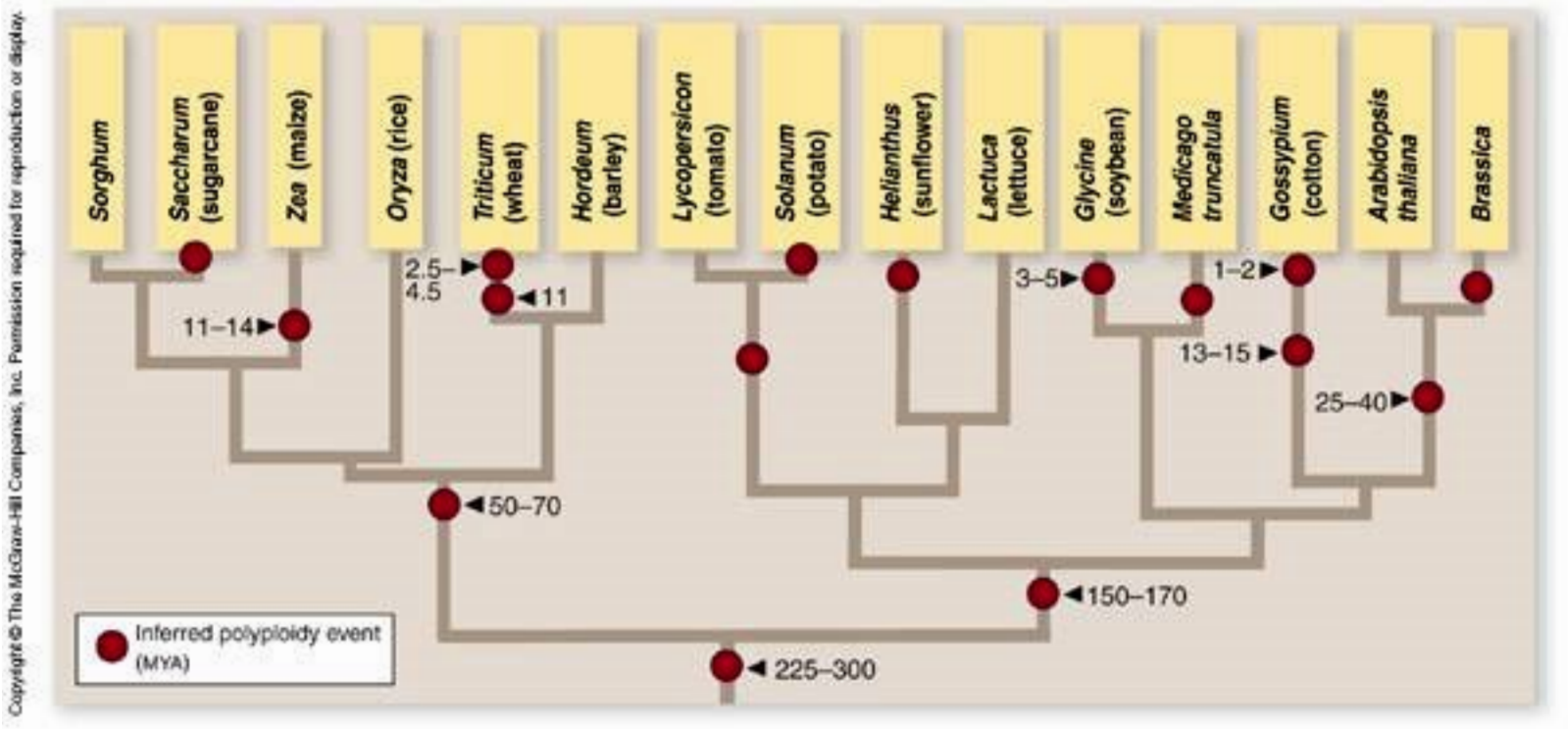


Figure 1 - Synoptic chart of cultivated wheats evolution: the diploid (2n = 14, AA) forms of *Triticum monococcum* (**a**) were naturally pollinated by weed species, possible *Aegilops speltoides* (2n = 14, BB?), in about 10,000 B.C. primitive farms. The subsequent genome duplication of hybrids by natural 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**).

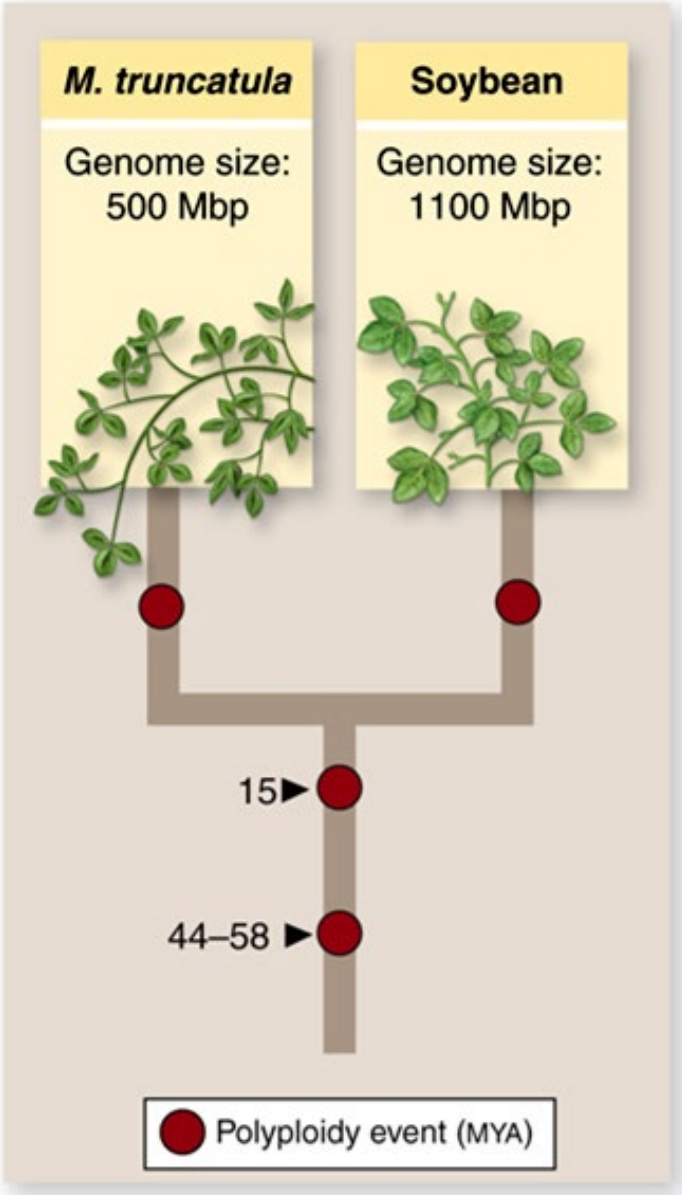
La poliploidia guida lo studio dell'evoluzione dei genomi

- **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

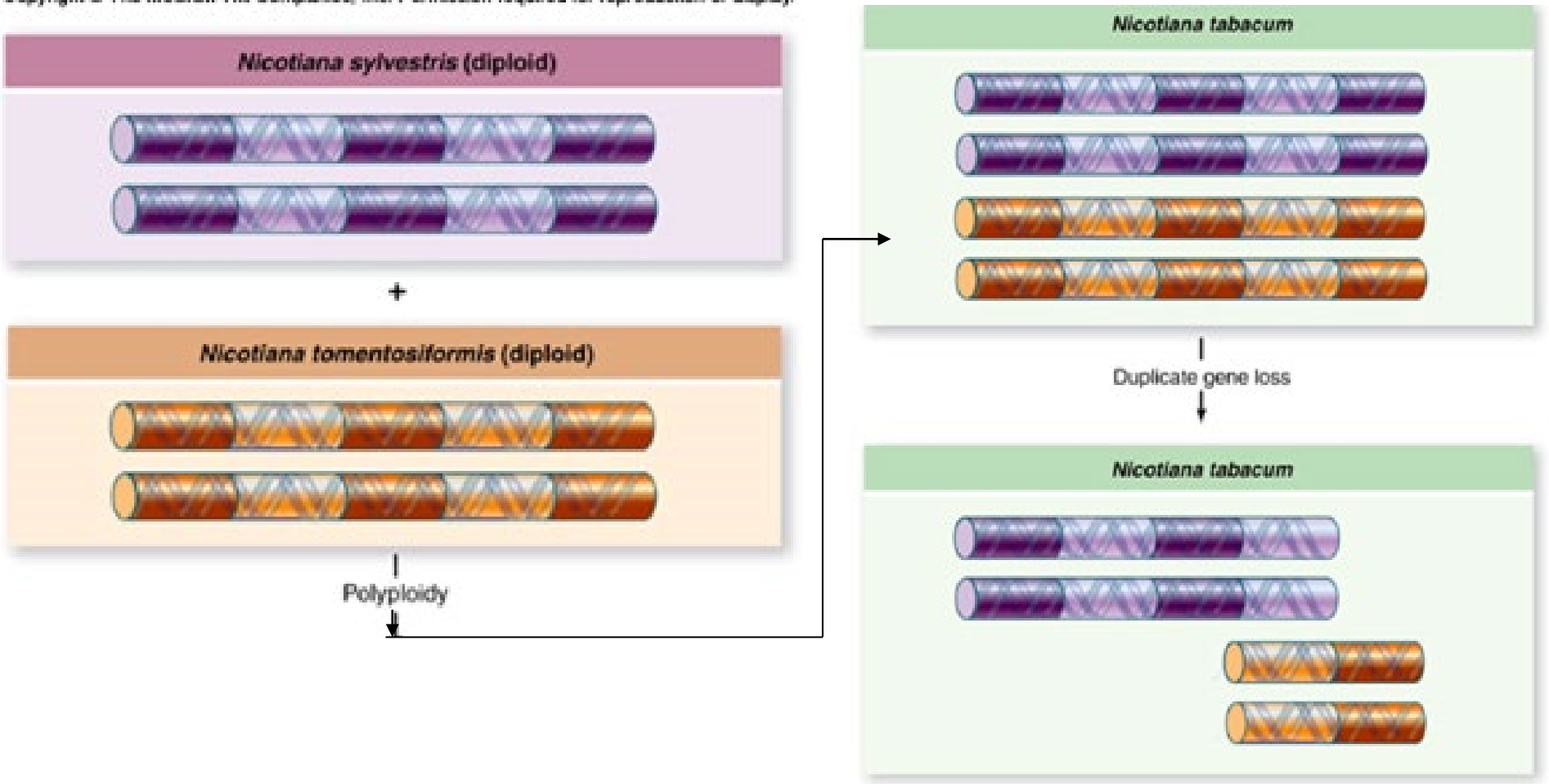


La poliploidia è diffusa nelle piante e ha avuto origini multiple durante l'evoluzione



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

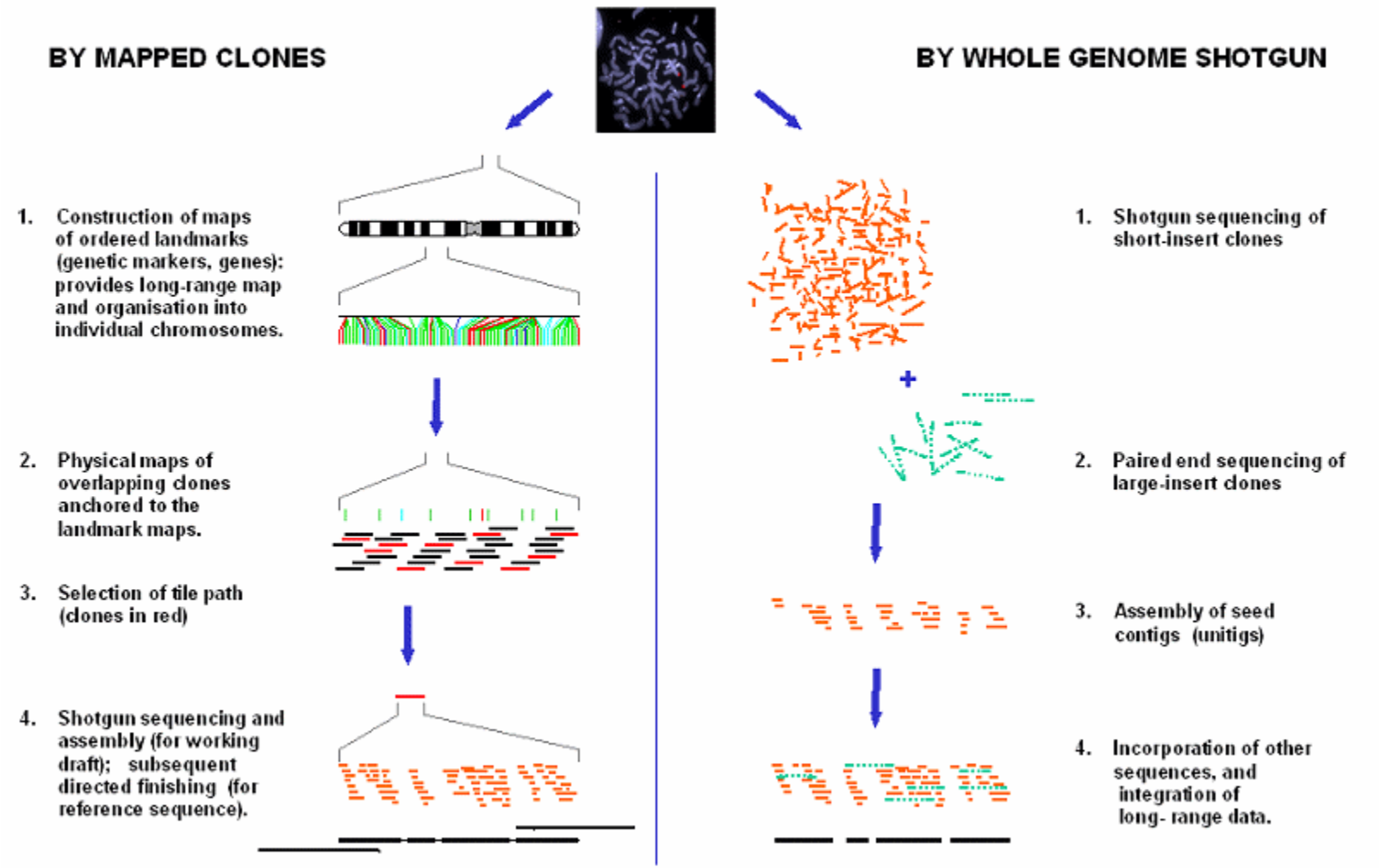
Molte specie autopoliploidi sono intolleranti all'INBREEDING, e hanno alti livelli di eterozigosità, importanti per la produttività

-> 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

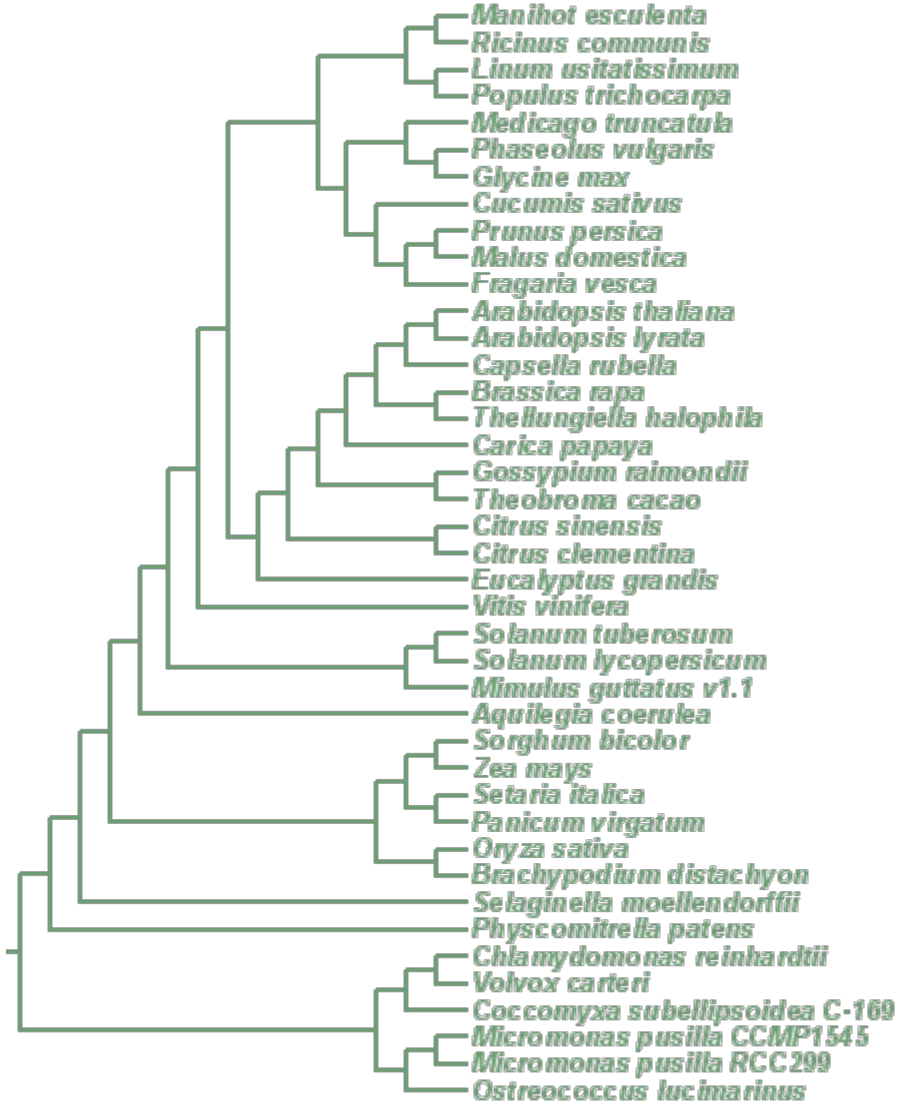
Creati sequenziando l'estremità 5' e/o 3' di mRNA isolati a caso e convertiti in cDNA (di solito 200–900 nt)

-> 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

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

Gene con funzione ignota

Trasferimento di annotazione

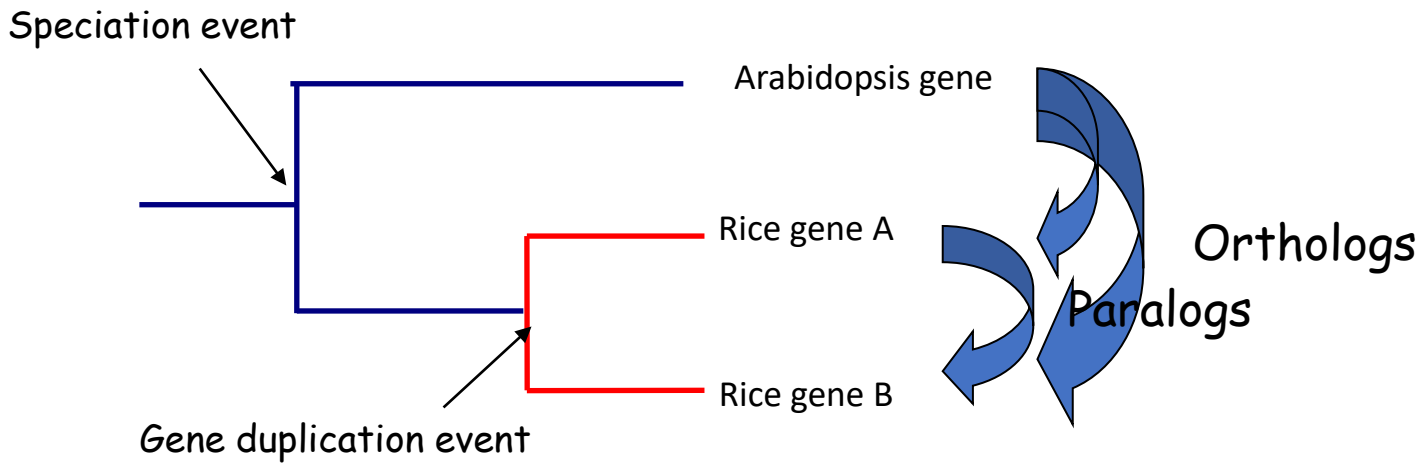
Specie modello

Geni omologhi

Gene con funzione X

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

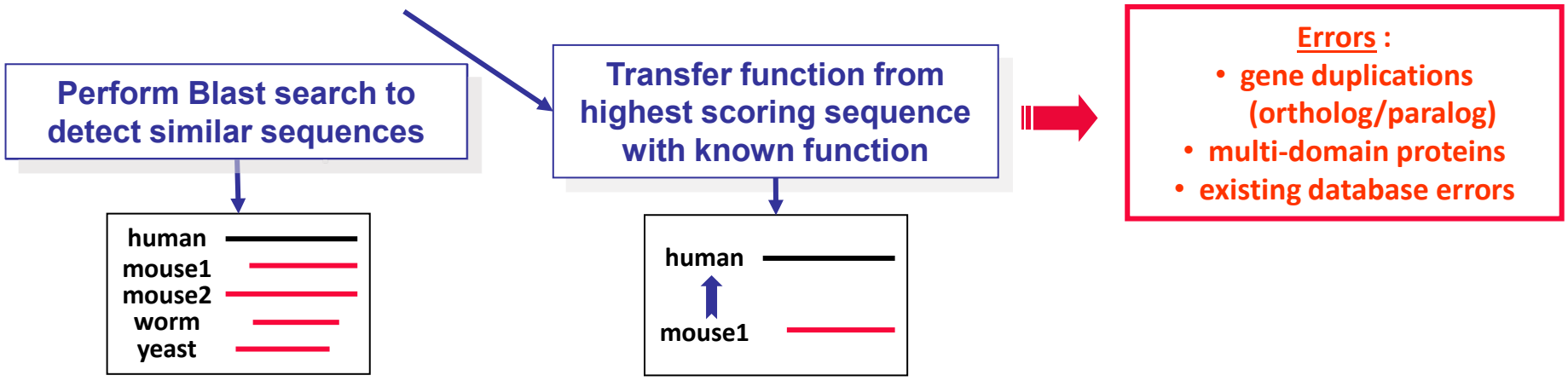
Come predirre l'omologia?

Similarità e omologia non sono la stessa cosa!

Geni simili si assomigliano sulla base di un'osservazione empirica

Geni omologhi 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?**

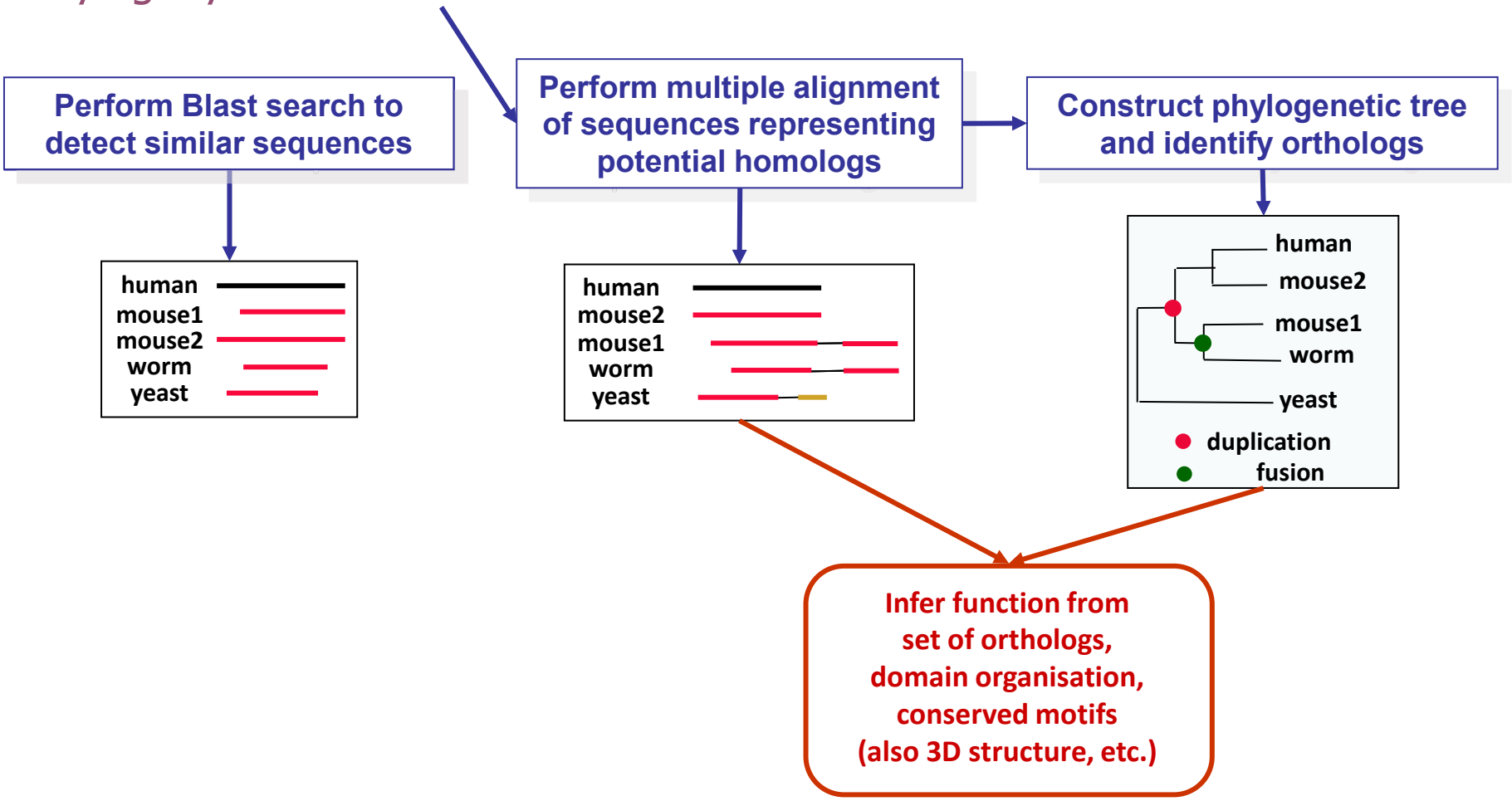
Due sequenze possono presentare similarità senza essere evolutivamente correlate!

- **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)

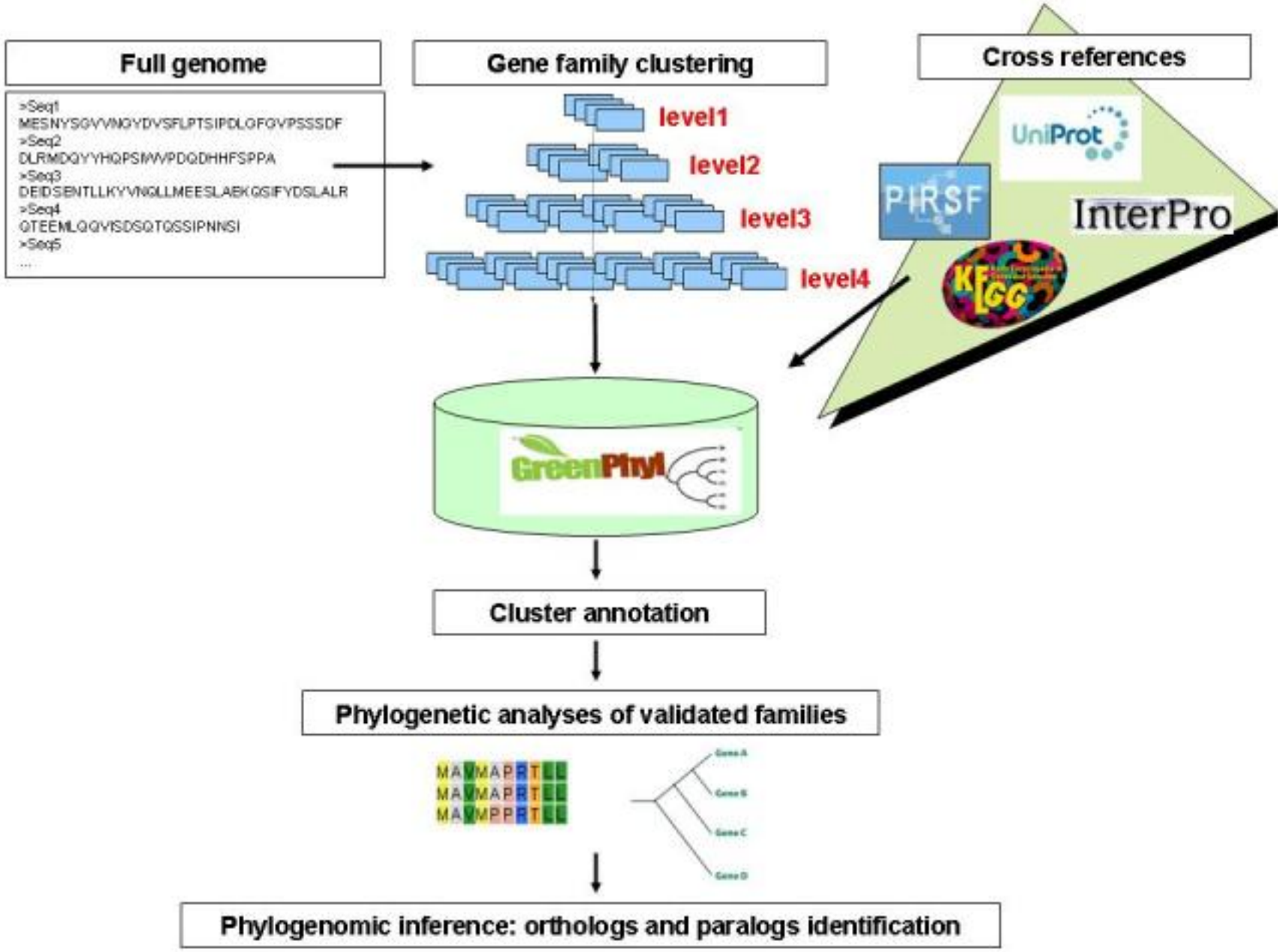
Due specie modello

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
MESNYSGVVNGYDVSFLPTSIPDLQFQVPSSSDF
>Seq2
DLRMDQYYHQPSMVVPDQDHHFSPPA
>Seq3
DEIDSENTLLKYVNGLLMEESLAEKGSIFYDSLALR
>Seq4
QTEEMLQQVISDSQTQSSIPNNSI
>Seq5
...
Gene family clustering
level1
level2
level3
level4
Cross references
UniProt
PIRSF
InterPro
KEGG
GreenPhyl
Cluster annotation
Phylogenetic analyses of validated families
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*	

Family_name (e.g. CRAS) 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 (85) 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 pl	Family	80	(653)	Y

Other InterPro signatures

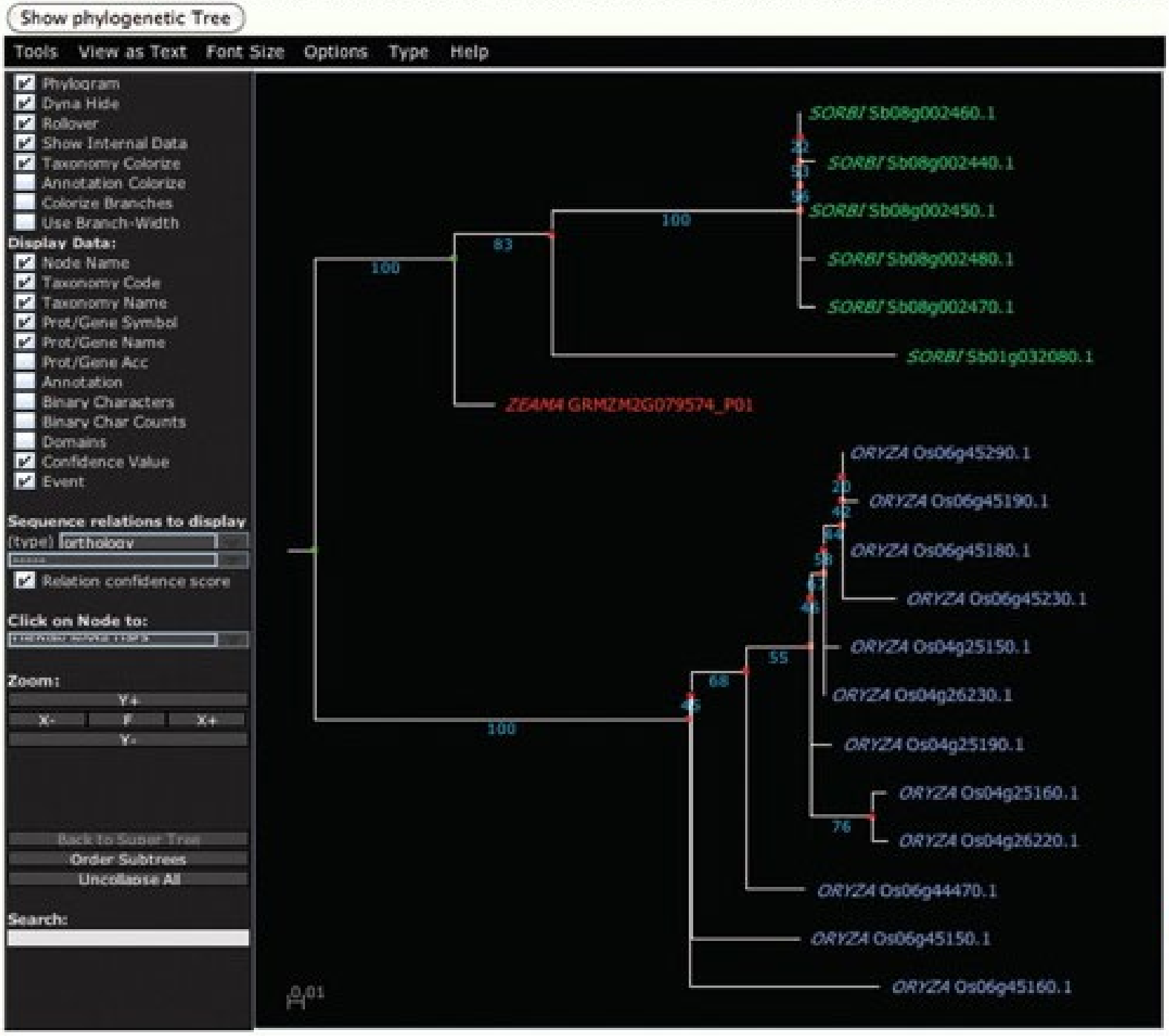
IPR	Annotation	Type	%	Occurence	Specificity
IPR014734	Pollen allergen, N-terminal	Domain	57	(466)	N
IPR009009	Barwin-related endoglucanase	Domain	58	(472)	N
IPR005132	Rare lipoprotein A	Domain	89	(722)	N
IPR007117	Pollen allergen/expansin, C-terminal	Domain	82	(667)	N
IPR007112	Expansin 45, endoglucanase-like	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-
Back to Super Tree
Order Subtrees
Uncollapse All
Search:
SORBI Sb08g002460.1
SORBI Sb08g002440.1
SORBI Sb08g002450.1
SORBI Sb08g002480.1
SORBI Sb08g002470.1
SORBI Sb01g032080.1
ZEAMA GRMZM2G079574_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
22
53
56
100
20
42
64
58
55
68
46
100
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)

Confronto:
Genomi delle piante e Genomi degli animali

I genomi delle piante contengono numerose classi di geni assenti o scarsamente rappresentati nei genomi animali.

I prodotti di questi geni “specifici” delle piante conprendono:

gli enzimi richiesti per la biosintesi della parete cellulare

Alcune proteine di trasporto, che spostano tra una cellula e l’altra nutrienti, composti tossici, metaboliti, proteine e ac. nucleici

Alcuni enzimi e altre macromolecole necessarie per la fotosintesi, come la Rubisco e le proteine di trasporto degli elettroni

I prodotti che sono coinvolti nel turgore cellulare e nelle risposte tipiche di un sistema di vita sessile, come il fototropismo ed il geotropismo

Numerosi enzimi e citocromi coinvolti nella produzione di centinaia di migliaia di metaboliti secondari delle piante in fioritura

Un numero molto elevato di geni R, legati alla resistenza ai patogeni e ai fattori associati

Il genoma dei vegetali condivide con quello degli animali molte famiglie geniche:

- coinvolte nella comunicazione intracellulare
- nella regolazione trascrizionale
- nella trasduzione dei segnali durante lo sviluppo

Altre famiglie di fattori di trascrizione sono tipiche delle piante.

composti a funzione ormonale esclusivi delle piante

PROGETTO GENOMA
Graminacee e Leguminose

Sono in corso progetti riguardanti il genoma di più di 60 specie di piante

Dal punto di vista economico, i più importanti di questi progetti riguardano le principali piante alimentari

Orzo

Mais

Miglio

Riso

Frumento

Lolium

Vite

Alfalfa

Soia

Fagiolo

Pomodoro

Patata

Melo

Cotone

Alcuni di questi genomi sono molto grandi (poliploidia e DNA ripetitivo)

ed il sequenziamento dell'intero genoma non è al momento realizzabile

Le ricerche si concentrano quindi sulla

genomica comparativa

Entrambi i genomi

Riso e mais hanno genomi relativamente piccoli e sono così importanti per le economie agricole dei Paesi Sviluppati che è stata data priorità al sequenziamento completo