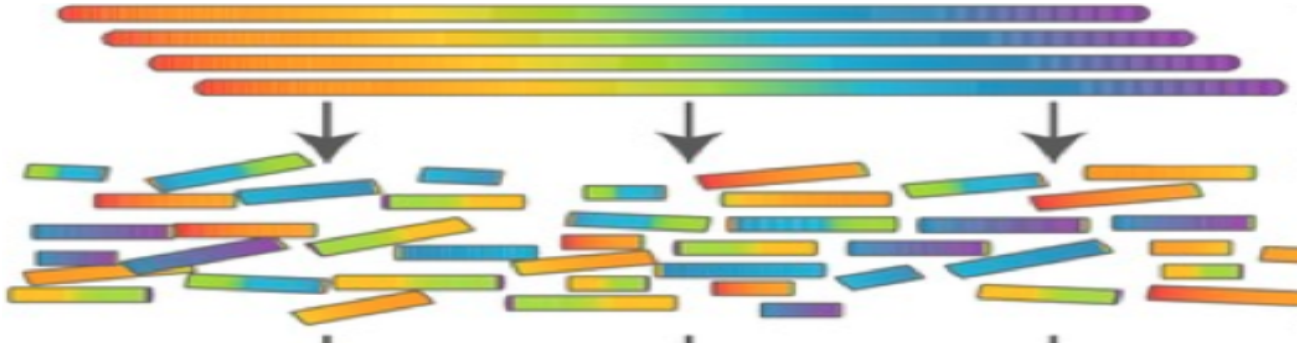
IL PUZZLE

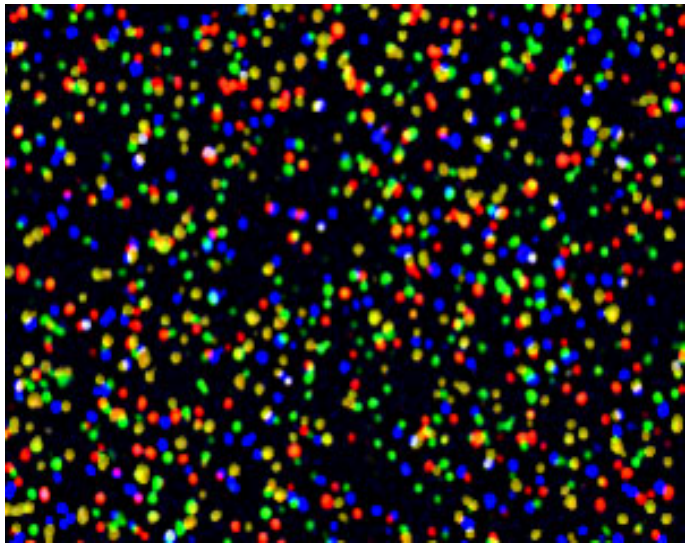


- Il DNA è una lunghissima sequenza di lettere ACTG.
  Per esempio 230.000.000 di lettere per il pesco, 690.000.000 di lettere per il melo
- Le attuali tecnologie (sequenziatori) non consentono di leggerle in sequenza, come faremmo con un libro
- Si può replicare la sequenza, si può spezzettare in maniera casuale la sequenza, se i pezzettini (read) sono sufficientemente corti si possono leggere i pezzettini
- Diciamo che la lunghezza di una read è circa 250 basi

Come ricomponiamo l'intera sequenza?

UN SEQUENZIATORE

Il sequenziatore



L'output

IL PUZZLE

E' come risolvere un Puzzle? Molto peggio!!!

- Si tratta di un Puzzle con più di 1.000.000 di pezzi
- Non abbiamo la scatola del Puzzle con sopra l'immagine da ricostruire
- Non abbiamo gli incastri tra i pezzi, ogni pezzo è un *rettangolo*

Sofia (6 anni): “**Non si può fare!!! Non c'è soluzione!**”

SEQUENZIAMENTO E ASSEMBLAGGIO

GACTTTGTAACATACAACCTTTAATCACG

ACATACAAC
CTTTAATCACG
GACTTTGTA

???

Determinare la più corta sequenza che contiene tutte le read.

Non si può! Qualsiasi sequenza ottenuta appiccicando le read va bene!

Mancano gli incastri!

SEQUENZIAMENTO E ASSEMBLAGGIO

SOLUZIONE: PIU' COPIE

GACTTTGTAACATACAACCTTTAATCACG

ACATACAAC
CTTTAATCACG
GACTTTGTA

GACTTTGTAACATACAACCTTTAATCACG

ATACAACCTT
GACTTTGTAAC
TAATCACG

ACATACAAC
CTTTAATCACG
GACTTTGTA
ATACAACCTT
GACTTTGTAAC
TAATCACG

INPUT

ACATACAAC
CTTTAATCACG
GACTTTGTA
ATACAACCTT
GACTTTGTAAC
TAATCACG

COMPUTAZIONE

GACTTTGTA
GACTTTGTAAC

GACTTTGTAACATACAAC
GACTTTGTAAC

GACTTTGTAACATACAAC
GACTTTGTAACATACAACCTT

GACTTTGTAACATACAACCTTTAATCACG
GACTTTGTAACATACAACCTT

GACTTTGTAACATACAACCTTTAATCACG
GACTTTGTAACATACAACCTTTAATCACG

SEQUENZIAMENTO E ASSEMBLAGGIO

Abbiamo reso facile una storia difficile:

- La stringa da ricostruire è lunghissima
- Ci sono errori sperimentali in tutte le fasi (e.g., non tutte le reads sono giuste)
- Sequenze ripetute potrebbero farmi *accorciare* la ricostruzione
- ….

Nonostante tutte queste difficoltà:

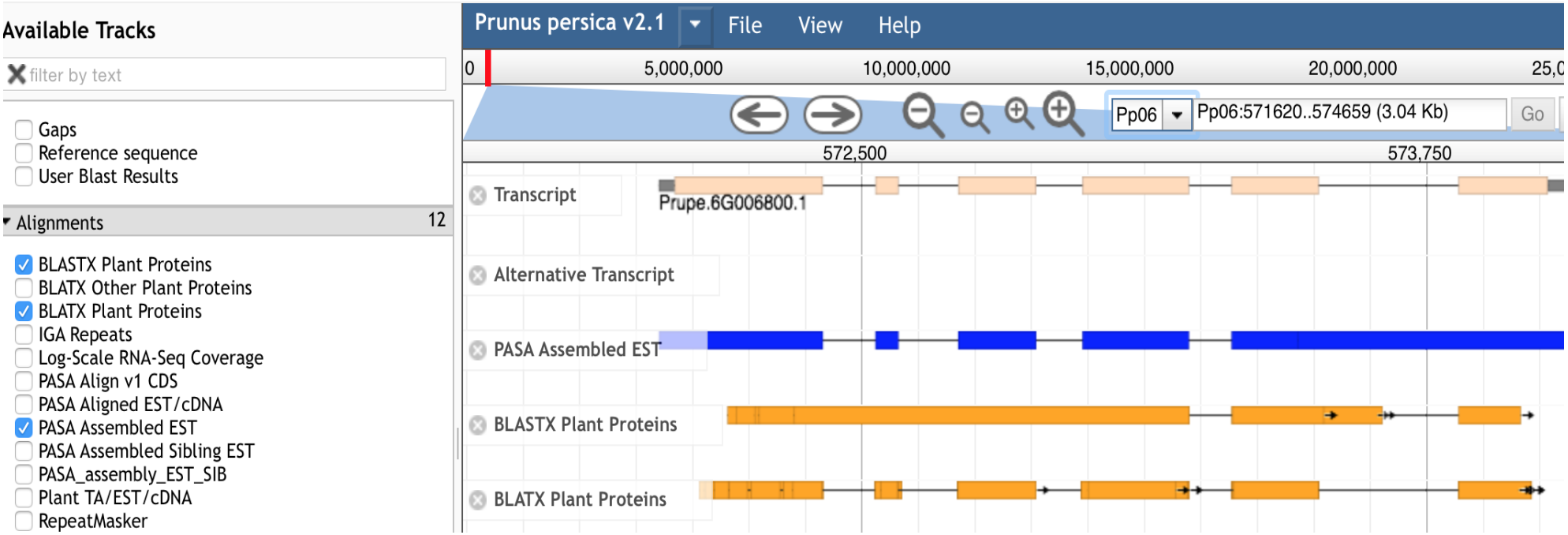
- Nel 2007 è stata sequenziata la Vite
- Nel 2013 è stato sequenziato il Pesco
- ...

UNIVERSITÀ DEGLI STUDI DI UDINE

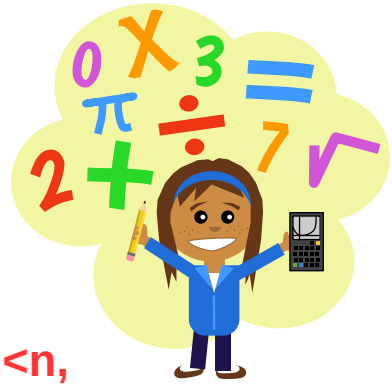
DATA BASE GENOMICI

JGI Phytozome 11
THE PLANT GENOMICS RESOURCE

JGI HOME



PER ESERCITARSI



Dato un testo T di lunghezza n e una parola P di lunghezza m<n,
Descrivere un algoritmo per trovare tutte le occorrenze di P in T

E' facile trovare una soluzione che nel caso peggiore fa n*m confronti tra caratteri

Riuscite a trovare una soluzione che nel caso peggiore fa circa n+m confronti?

E se doveste anche gestire degli errori (inserimenti e cancellazioni di caratteri in T)?