Alcuni aspetti legati al calcolo bioinformatico su CRESCO

Giuseppe Aprea
UTMEA-CAL

Principali attività bioinformatiche ENEA legate al calcolo

- Assemblaggio de Novo*
- Trascrittomica
- Analisi filogenetica
- Metagenomica*
- Analisi delle sequenze

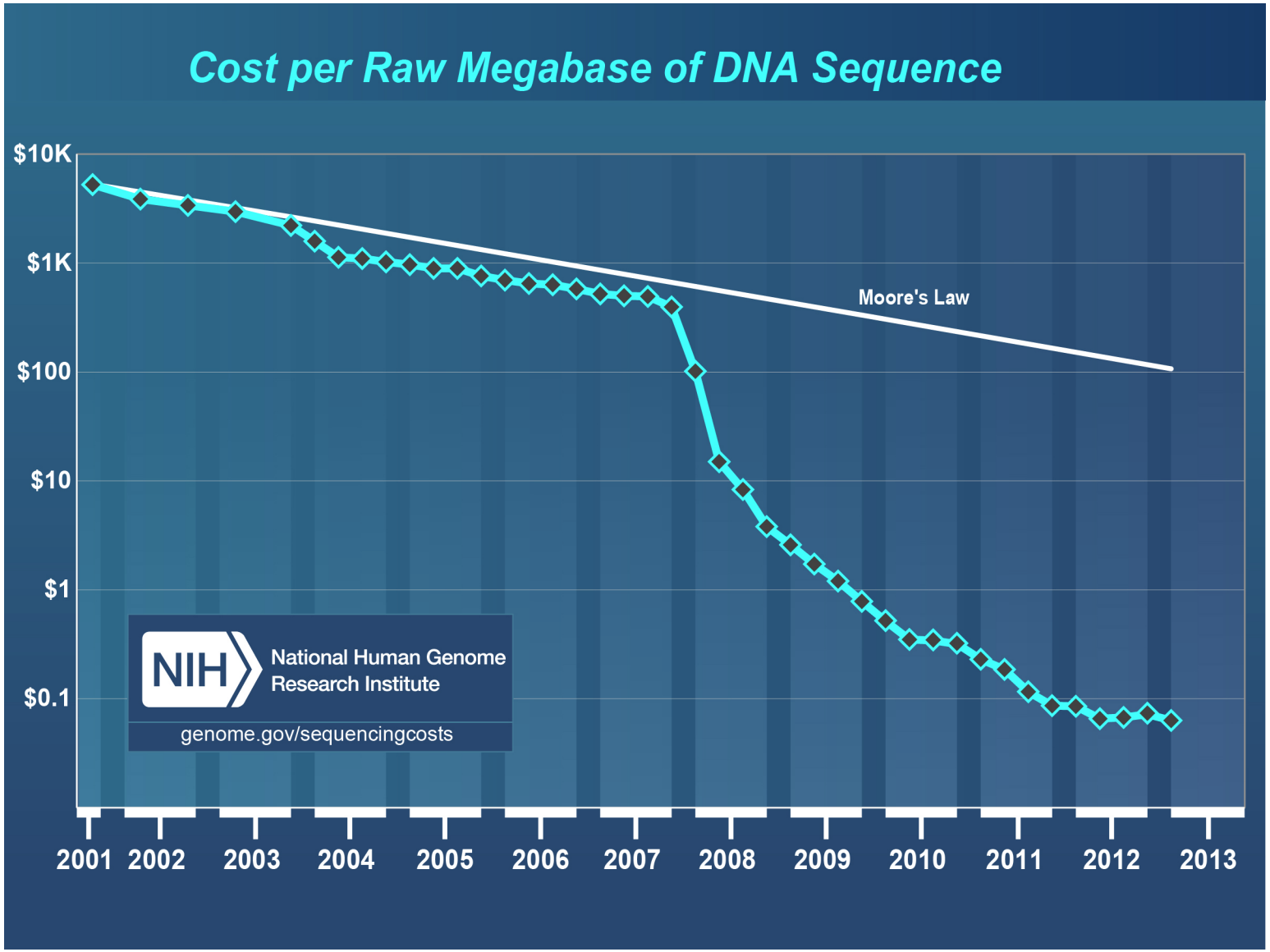
- In tutti questi casi l'input è costituito da dati di sequenziamento.

L'evoluzione dei dati (1)

Technology	First Generation (Sanger) Sequencing	Next Generation Sequencing			
		Roche 454 (GS FLX Titanium XL+)	Illumina (HiSeq 2000 dual flow cell)	Abi Solid (Solid 4)	Ion Torrent (PGM)
Bases per RUN	~1Mb	700Mb	600Gb	100Gb	1Gb
Time per RUN	~1 day	~1 day	~11 days	~14 days	4.5h
Reads per RUN	NA	1 million	6 billions (paired-end) or 3 billions (single)	1.4 billions	millions
Reads length	up to 1000 bp	up to 1000 bp (mode 700 bp)	2*100 bp	2*50 bp	35-400 bp

- la lunghezza delle sequenze si è accorciata di un fattore 10 circa (eccetto il taso della tecnologia 454)
- Nel sequenziamnto di seconda generazione il throughput è aumentato enormemente (fino a un fattore ~10^5)
- Al giorno d'oggi la tecnologia più affermata è Illumina di conseguenza si incontrano molto spesso reads corte.

L'evoluzione dei dati (2)



L'evoluzione dei dati (3)

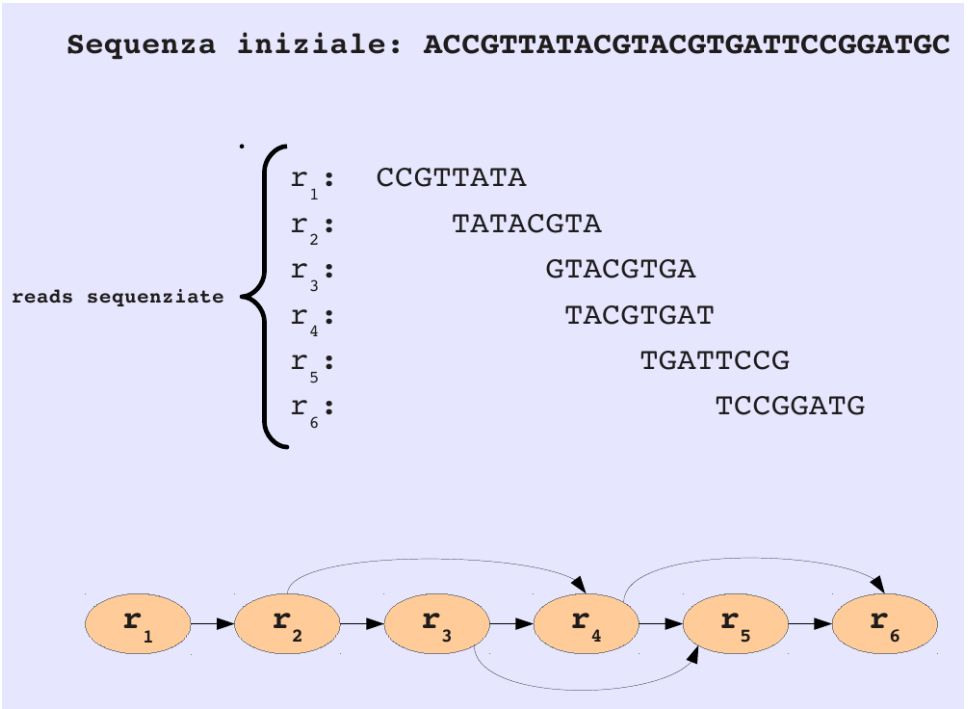
Il prezzo da pagare in seguito agli sviluppi delle tecniche di sequenziamento è dato da:

- L'analisi dati è diventata molto più complessa
- È aumentata la richiesta di risorse di calcolo

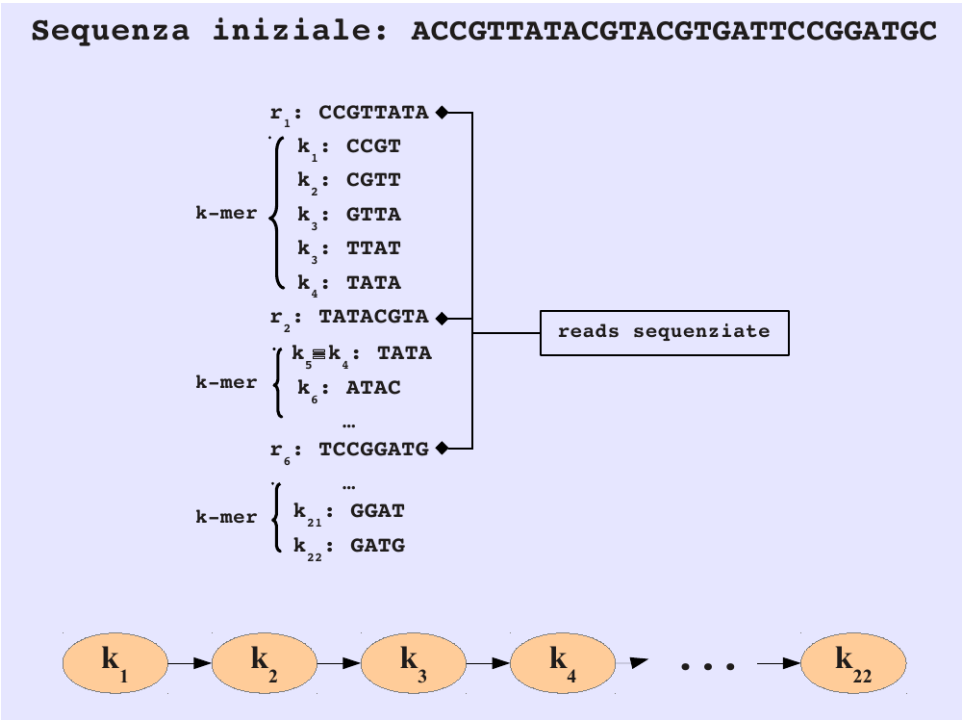
Assemblaggio: l'evoluzione degli algoritmi (1)

Gli algoritmi di assemblaggio più usati fanno uso di strutture a grafo di 2 tipi:

Grafi di overlap



Grafi di de Bruijn



Assemblaggio: l'evoluzione degli algoritmi (2)

- Gli algoritmi basati su grafi di overlap sono più indicati per reads medio-lunghe (454)
- Gli algoritmi basati su grafi di Bruijn sono più indicati per reads corte e di alta qualità (Illumina)
- Su CRESCO sono attualmente installati sia software basati sul primo tipo di algoritmo (Newbler) che sul secondo (Abyss, Soap2, Velvet)

Assemblaggio: memory footprint

Algoritmi basati su grafi di de Bruijn (Abyss):

Memoria totale = numero_di_kmer_unici * byte_per_kmer

numero_di_kmer_unici ≈ [genome_size] + [numero_reads * (l-k+1) * p]

l=lunghezza read, k=lunghezza kmer (NB:l>k), p=probabilità che ci sia almeno un errore per read

byte_per_kmer = 8 + maxk/4

maxk = 32, byte_per_kmer=16; maxk = 64, byte_per_kmer=24; maxk = 96, byte_per_kmer=32.

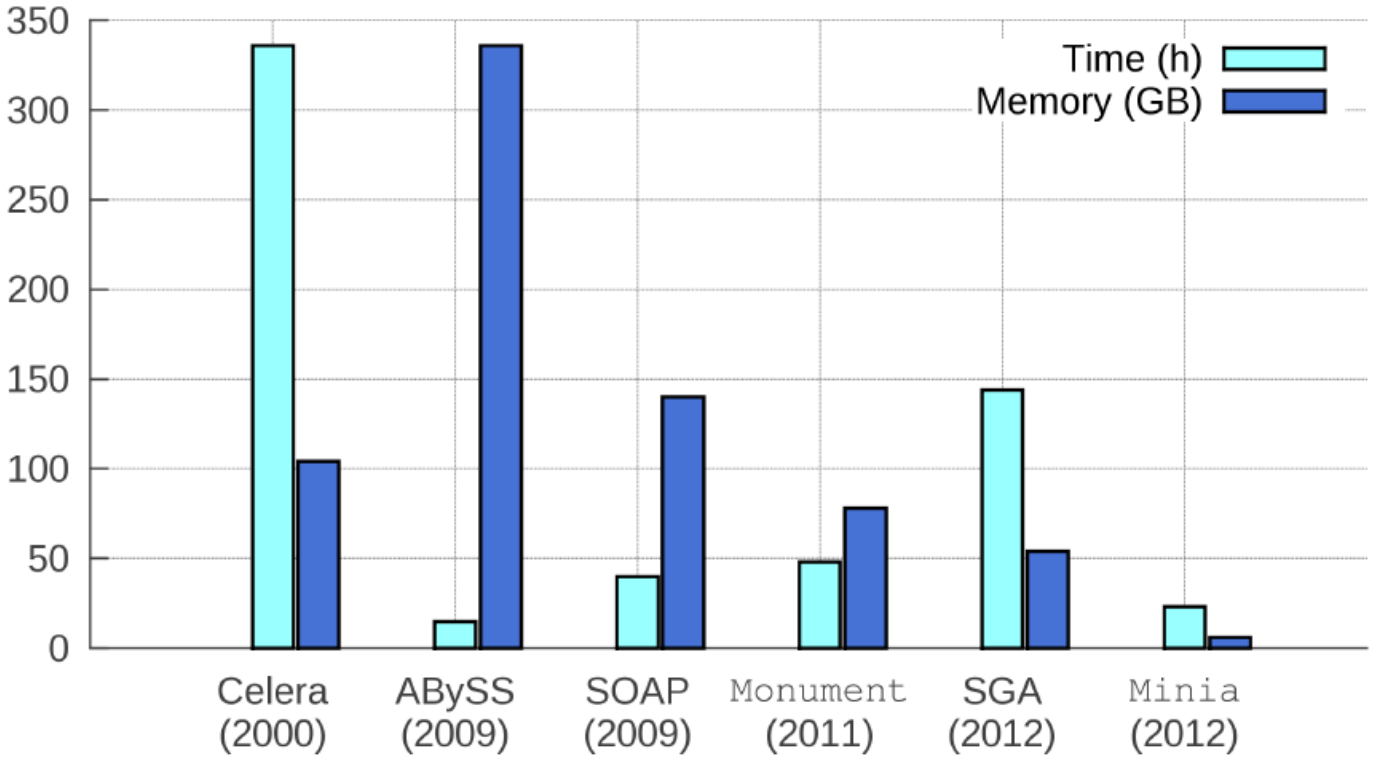
Es: genome_size=1 Gbase, coverage 20x, l=100bp, k=64, maxk = 64, p=0.63

Memoria totale ≈ 135 GB

Assemblaggio: esempi reali

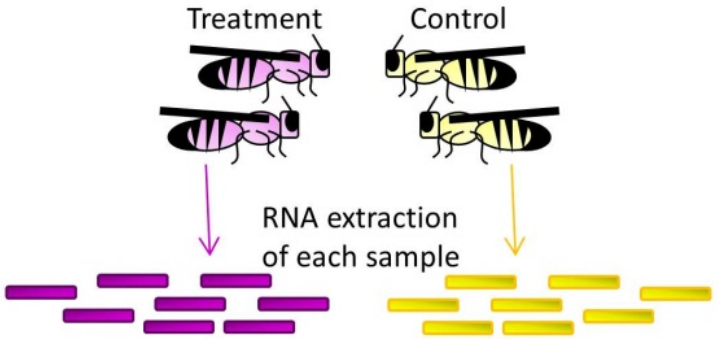
- AbySS: Assemblaggio del genoma umano in 87 h su cluster di 21 nodi da 8 core, ciascuno con 16 GB of RAM (totale di 168 core, 336 GB RAM)[Simpson et al. 2009].
- SOAPdenovo: assemblaggio del genoma umano in 40 h su singolo nodo con 32 core e 512 GB di RAM [Li et al. 2010].

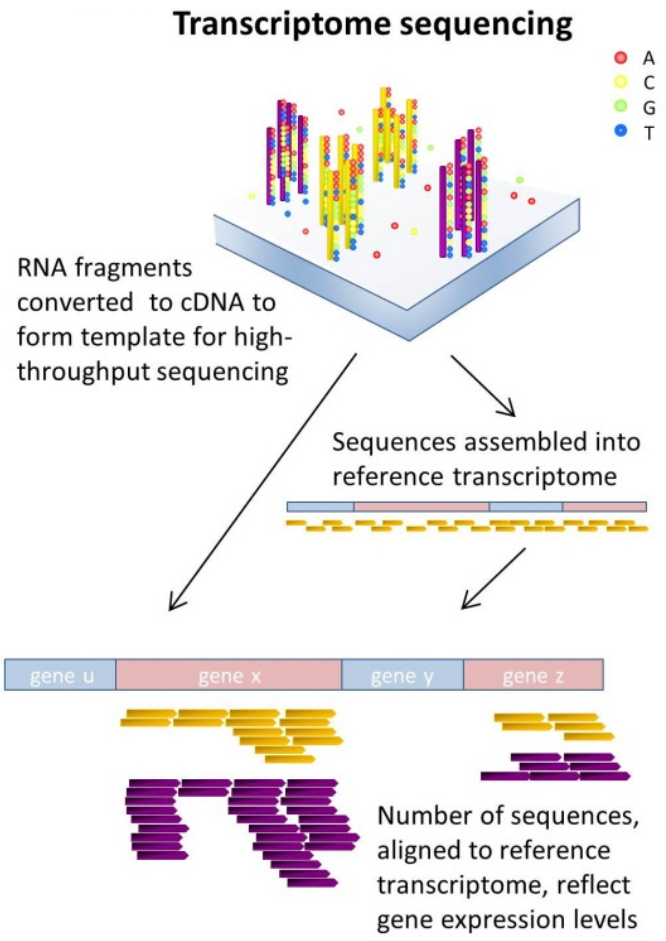
Ultra low memory assembly



Approccio molto promettente ma ancora non consolidato.

RNA-seq (1)





Gene ID	Transcript abundance in Treatment sample	Transcript abundance in Control sample
ABc00001	3778	3894
ABc00002	768	189
ABc00003	1087	1476

RNA-seq (2)

É stata implemetata su CRESCO una pipeline per la determinazione dell'espressione differenziale dei trascritti:

- tophat: allinemento delle sequenze al genoma;
- cufflinks,cuffmerge,cuffdiff: calcolo dei livelli d'espressione, identificazione di eventuali nuovi trascritti o varianti di splicing.

- La pipeline è in grado di processare qualche decina di campioni con genoma contenete alcune decine di geni in 2-3 giorni.
- Gira principalmente su code a basso parallelismo.

Analisi filogenetica genome-wide

É stata implemetata su CRESCO una pipeline per la determinazione delle sequenze consenso dei trascritti genici e per il loro confronto:

- Bowtie2: allinemento delle sequenze al riferimento
- Samtools: determinazione delle sequenze consenso
- Clustalw: riallinemento tra i consensi di uno stesso trascritto
- PAML(yn00): analisi della sostituzioni sinonime e non sinonime.

- La pipeline è in grado di processare un insieme di qualche decina di migliaia di geni per 4-5 campioni in circa un paio di giorni.
- Gira principalmente su CRESCO1

Conclusioni

Le risorse di calcolo attuali e quelle prospettate per CRESCO4 sono adeguate all'esecuzione della maggior parte dei calcoli bioinformatici descritti. Tuttavia esistono alcune criticità:

- ci sono solo 2 “nodi” con memoria ad immagine unica dell'ordine di 100GB o superiore. Si tratta di 2 nodi un pò datati. Non vi sono nodi con memoria maggiore di 256GB, necessari per l'assemblaggio di genomi complessi di grandi dimensioni.

- in alcuni casi l'analisi di grandissime quantità di sequenze può richiedere tempi superiori ai massimi consentiti dalle code attuali (10 giorni). I codici coinvolti tipicamente non consentono resume (o non lo consentono in maniera semplice) e non lavorano in parallelo distribuito (Newbler, ClustalOmega).

- Lo spazio disco necessario per i dati grezzi e l'area di lavoro è dell'ordine di grandezza di 10TB e aumenta di circa 1TB all'anno (la messa in esercizio di cresco3 aiuta già molto da questo punto di vista).