

Implementazione, creazione e ottimizzazione di una pipeline per l'analisi biofisica su cluster a basso consumo energetico

Daniele Dall'Olio

Relatore: Dott. Enrico Giampieri

Correlatori: Prof. Gastone Castellani Ing. Andrea Ferraro

ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

22 Settembre 2017

Problema

- Costo medio elevato
- Consumo energetico elevato
- Spese per il raffreddamento elevate

Conseguenze

- Minor accessibilità
- Poche unità acquistabili
- Ridotta scalabilità e flessibilità per aggiornare l'hardware dei server

Tecnologia di calcolo low power

Vantaggi

- Costo delle singole unità basso
- Consumo elettrico inferiore
- Flessibilità nell'acquisto di nuovi hardware

Svantaggi

- Cache ridotta
- Potenza inferiore
- Numero inferiori di core

Obiettivo

Ottenere risultati comparabili a quelli ottenuti con i nodi tradizionali.

<i>Nodo</i>	<i>CPU</i>	<i>Memory</i>	<i>Storage</i>	<i>Costo*</i>	<i>Consumo*</i>
<i>xeond</i>	1x Xeon D-1540	16 GB	8 TB(HDD)	€1000	60 W
<i>avoton</i>	1x Atom C2750	16 GB	5 TB(HDD)	€600	30 W
<i>n3700</i>	1x Pentium N3700	8 GB	0.5 TB(SSD)	€130	8 W
<i>bio8</i>	2x Xeon E5-2620v4	128 GB	2 TB(HDD)	€10000	180 W

* I valori di costo e consumo energetico sono stimati.

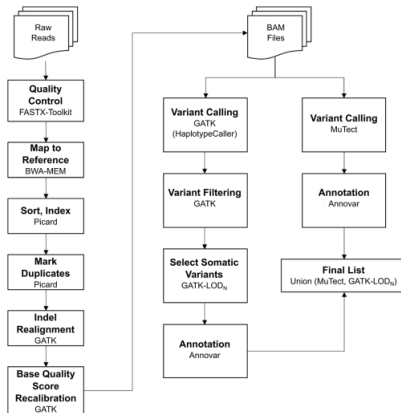
Tabella: Caratteristiche dei nodi.

<i>CPU</i>	<i>Microarchitecture(Platform)/litho</i>	<i>Freq(GHz)</i>	<i>Cores</i>	<i>Cache</i>	<i>TDP</i>
Xeon D-1540	Broadwell/14nm	2.0(2.60)	8(16)	12 MB	45 W
Atom C2750	Silvermont(Avoton)/22nm	2.40(2.60)	8	4 MB	25 W
Pentium N3700	Airmont(Braswell)/14nm	1.60(2.40)	4	2 MB	6 W
Xeon E5-2620v4	Broadwell – EP/14nm	2.10(3.00)	8(16)	20 MB	85 W

Tabella: Caratteristiche delle CPU.

GATK-LODn

Requisiti molto elevati in termini di potenza di calcolo, di occupazione di memoria e di spazio d'archiviazione.

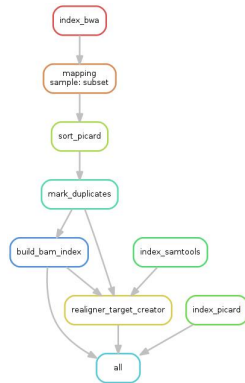


NGS

- Succede al HGP e supera il metodo Sanger
- Comprende le nuove tecniche per il sequenziamento del DNA
- Tecniche più rapide e meno costose
- Utilizzo della Teoria dei Network
- **Shotgun Sequencing**

Struttura delle simulazioni

Una parte di GATK-LOD_n è stata reimplementata nel tool Snakemake.



Regole

Indipendenti dal paziente

- Indicizzazione per BWA
- Indicizzazione per Picard
- Indicizzazione per Samtools(e GATK)

Dipendenti dal paziente

- **Mapping**: mappatura delle sequenze del paziente sul riferimento(in SAM).
- **Sort Picard**: riordinamento dei file SAM(in BAM).
- **Mark Duplicates**: identificazione dei duplicati.
- **Build BAM**: indicizza il file BAM per velocizzare l'analisi.
- **Realigner**: determina gli intervalli che necessitano probabilmente del riallineamento Indel.

Analisi effettuate

- Tempo di esecuzione
- Memoria utilizzata

Simulazioni effettuate

numero di letture	dimensione su disco
1×10^5	2x 28.4 MB
1×10^6	2x 284.9 MB
3×10^6	2x 854.9 MB
9×10^6	2x 2.6 GB
4.5×10^7	2x 12.8 GB

Tabella: Stima della dimensione dei subset in relazione al numero di letture. L'ultimo valore si riferisce all'intero paziente.

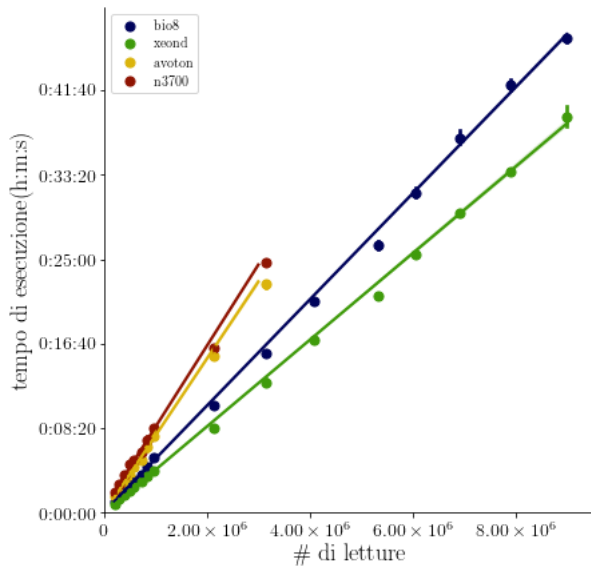


Figura: Tempi per Mapping.

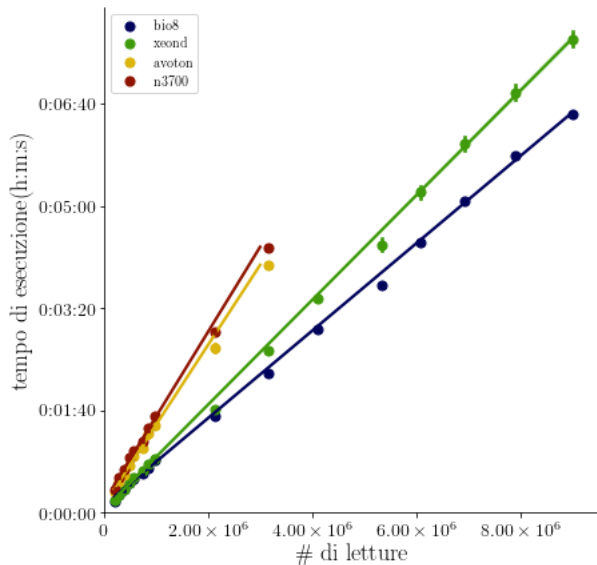


Figura: Tempi per Sort Picard.

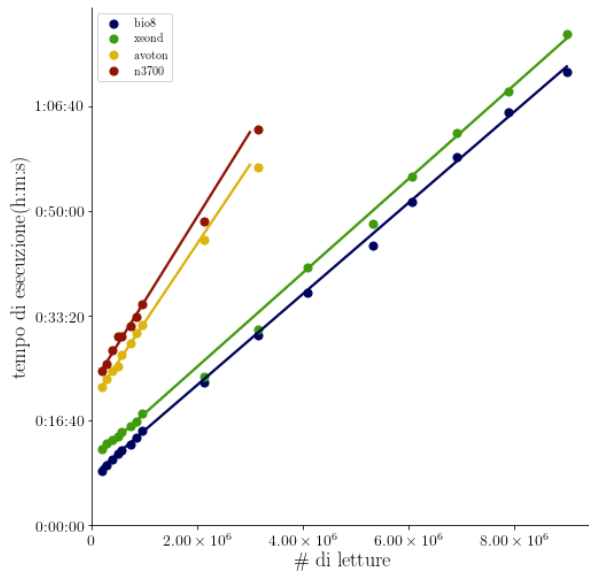


Figura: Tempi complessivi.

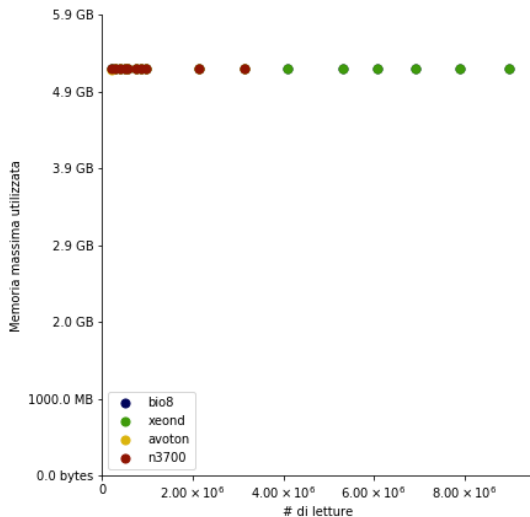


Figura: Mapping.

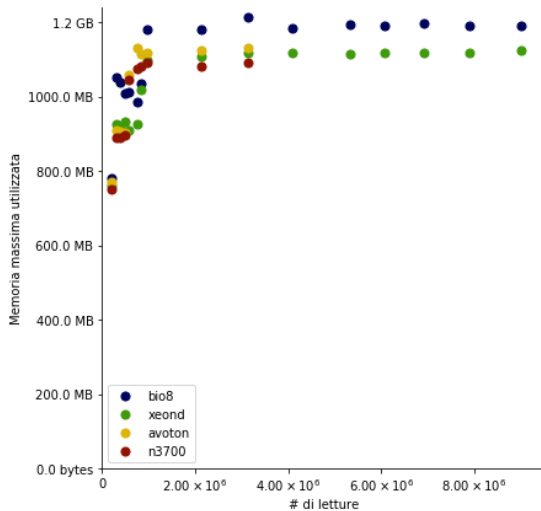


Figura: Sort Picard.

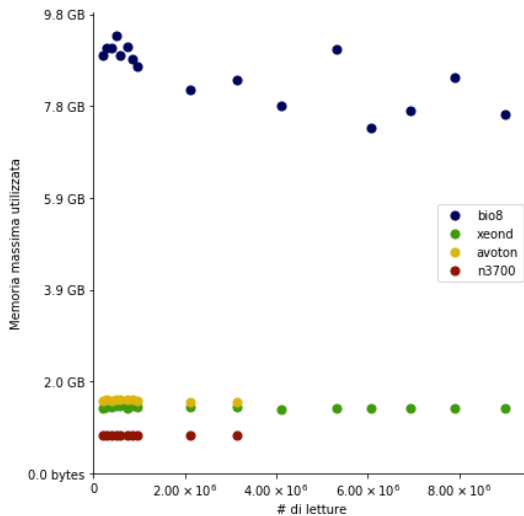


Figura: Realigner.

Tempo di esecuzione

- avoton e n3700 impiegano il doppio del tempo
- xeon è comparabile a bio8 consumando un terzo dell'energia e costando 10 volte di meno

Memoria utilizzata

- Saturazione
- Adattamento dinamico
- Sempre inferiore al massimo di memoria accessibile

Conclusione

In base a questi risultati questa pipeline di calcolo bioinformatico sembra essere realisticamente eseguibile anche su nodi a bassa potenza senza una perdita considerevole di prestazioni.

Sviluppo futuro

- Simulazioni a core multipli sui singoli nodi
- Completamento della pipeline
- Simulazioni su cluster

Pubblicazione

Cercheremo di completare il progetto e infine di pubblicarlo.