Feedback dalla comunità di bioinformatica

Ernesto Picardi

University of Bari
IBIOM-CNR

Incontro annuale degli utenti ReCaS-Bari - 4 luglio 2017

Next-Generation Sequencing

La nuove tecnologia di sequenziamento (NGS) hanno rivoluzionato in modo significativo il mondo della bioinformatica e della biologia computazionale con la produzione massiva di dati genomici. La bioinformativa moderna, quindi, offre ulteriori sfide come la condivisione, l'archiviazione, l'integrazione e l'analisi di questi BIG DATA.

NextSeq Series

HiSeq Series

HiSeq X Series†

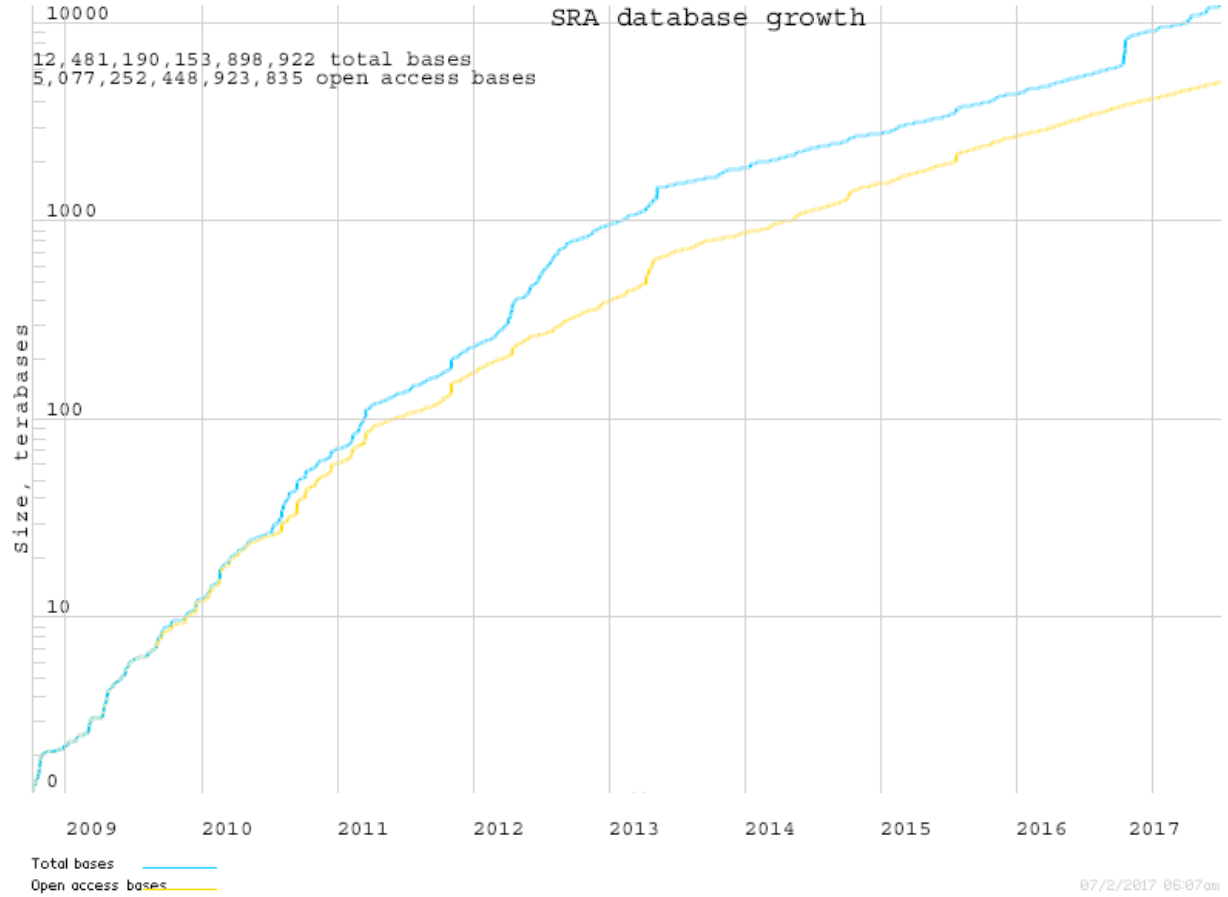
NovaSeq Series

Run Time	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	< 3 days	19–40 hours‡
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb§
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion|
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

www.illumina.com

Next-Generation Sequencing

L'elevata mole di dati prodotti dalle tecnologie NGS è testimoniata dalle dimensioni di uno dei più importanti database pubblici per NGS, lo Sequence Read Archive (SRA) presso NCBI (nato solo nel 2009).



Piattaforme NGS presso IBIOM/UNIBA

Illumina MiSeq
- > 50 M reads/run
- Paired 2x300
- Max output 15 GB

Illumina NextSeq 500
- > 800 M reads/run
- Paired 2x150
- Max output 250 GB

- **DNA sequencing (DNA-Seq)**
  - ri-sequenziamento genomico (SNPs, CNV, GWAS)
  - sequenziamento de novo
  - identificazione di varianti strutturali genomiche (cancer genome)
  - Epigenomica (stato della cromatina and metilazione del genoma)
  - Metagenomica (analisi del microbiota di cambioni clinici e ambientali)

- **RNA sequencing (RNA-Seq)**
  - Analiai qualitativa e quantitativa del trascrittoma
  - Identificazione e caratterizzazione di miRNA e altri ncRNA
  - RNA editing
  - Metatrascrittomica (analisi funzionale di campioni ambientali)

Risorse di calcolo presso IBIOM/UNIBA

3 Xserve Apple
- 16x3 GB RAM
- RAID 3 TB

HP Proliant Server
- 256 GB RAM
- 40 cores
- 2 RAIDs (24 + 36 TB)

HP Proliant Server
- 2 TB RAM
- 80 cores
- 2 RAIDs (36 + 48 TB)
- 4 GPUs

Attività Bioinformatiche attive presso il ReCaS

- Analisi del trascrittoma eucariotico con focus su splicing alternativo e RNA editing (tool usati: REDItools, SAMtools, Blat, MATS)
- Utilizzo di dati NGS depositati in archivi pubblici come SRA (Sequence Read Archive) o in *repository* specializzati come TCGA (tool usati: Aspera connect, sratoolkit, tcga client)
- Utilizzo di software per l'allineamento dati genomici, trascrittomici e metagenomici (tool usati: GSNAP, bwa, STAR, HiSat, Blat)
- Analisi di dati di Meta-barcoding attraverso l'applicazione della pipeline BioMaS (Bioinformatic analysis of Metagenomics amplicons)
- Caratterizzazione del genoma mitocondriale umano (tool usati: MToolBox, GSNAP, SAMtools, muscle, picard tools)

Statistiche delle attività presso il ReCaS

- Sottosistema utilizzato: Sistema Batch; Sistema HPC; Portale ReCaS Science Gateway
- Il sistema viene utilizzato: Quotidianamente
- Il sistema sarà utilizzato ancora per più di un anno
- Numero di job sottomessi: da 100 a 1000 al giorno quando utilizzato
- Storage utilizzato: 100TB (ma non sufficiente !!!)

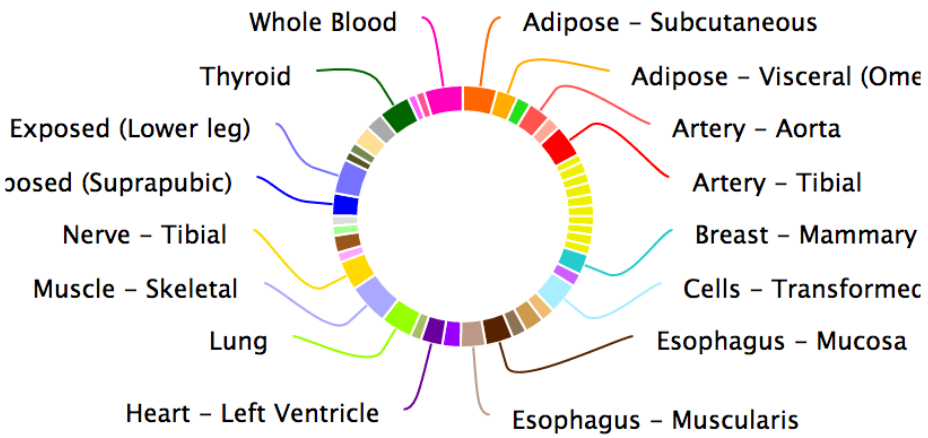
Il sistema HPC è utilizzato principalmente per lanciare i programmi di allineamento che, nonostante non lavorino in parallelo, richiedono molta RAM (in media 50-60GB con punte superiori a 110GB).

Il sistema batch è utilizzato per i programmi che richiedono poca RAM e per completare le analisi in cui è necessario lanciare molti job.

Use Case: RNA editing nei tessuti umani

L'RNA editing è un processo co-/post-trascrizionale in cui gli RNA primari sono modificati in specifiche posizioni mediante inserzioni, delezioni o sostituzioni di basi.
L'RNA editing è pervasivo nell'uomo dove svolge diverse funzioni (molte ancora da definire) e la sua deregolazione è stata associata a molte patologie neurologiche e neurodegenerative.
Da qualche anno abbiamo iniziato a caratterizzare gli eventi di RNA editing nei diversi tessuti umani sfruttando i dati trascrittomici prodotti dal consorzio GTEx.

V6p Release	# Tissues	# Donors	# Samples
Total	53	544	8555
With Genotype	53	449	7333
Has eQTL Analysis*	44	449	7051

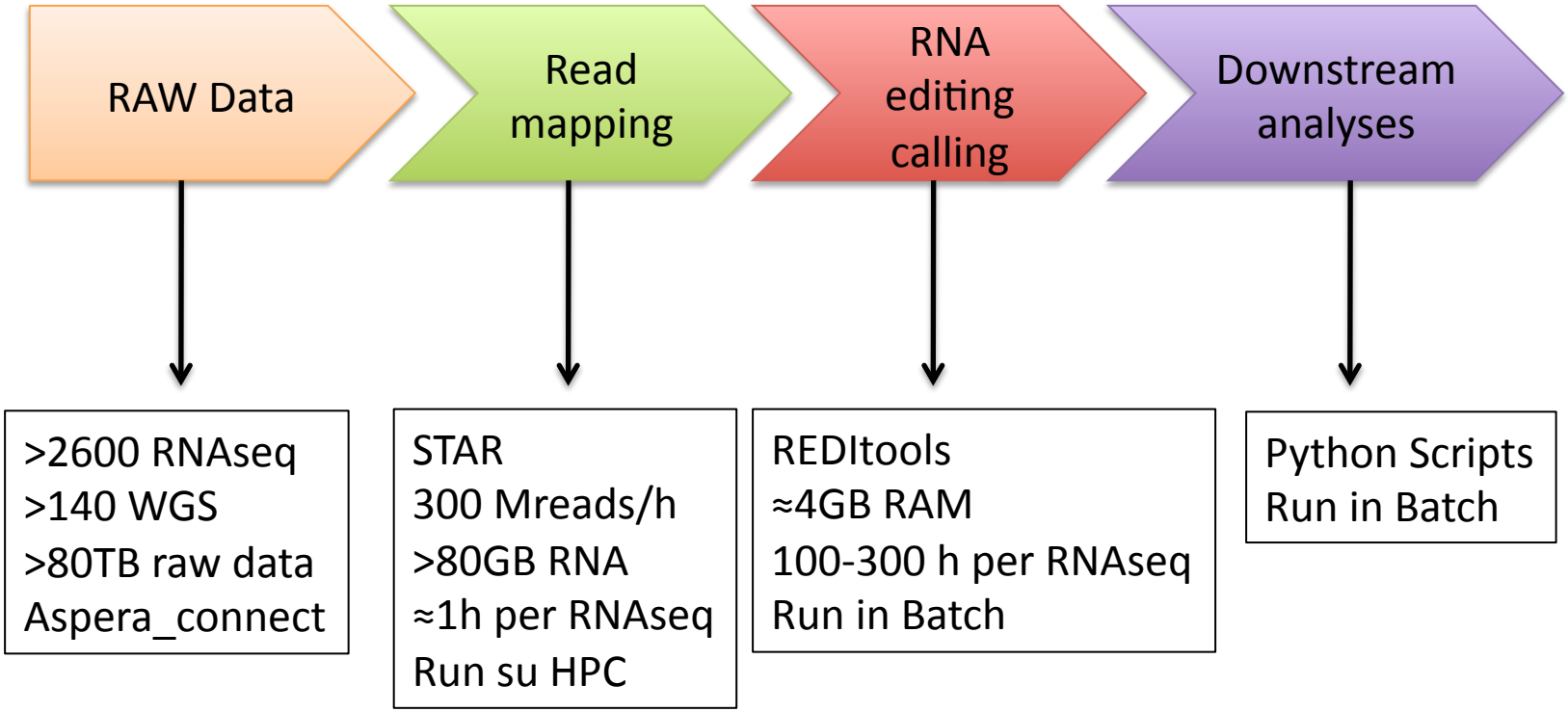


GTEx Portal

ReCaS Bari

Use Case: RNA editing nei tessuti umani

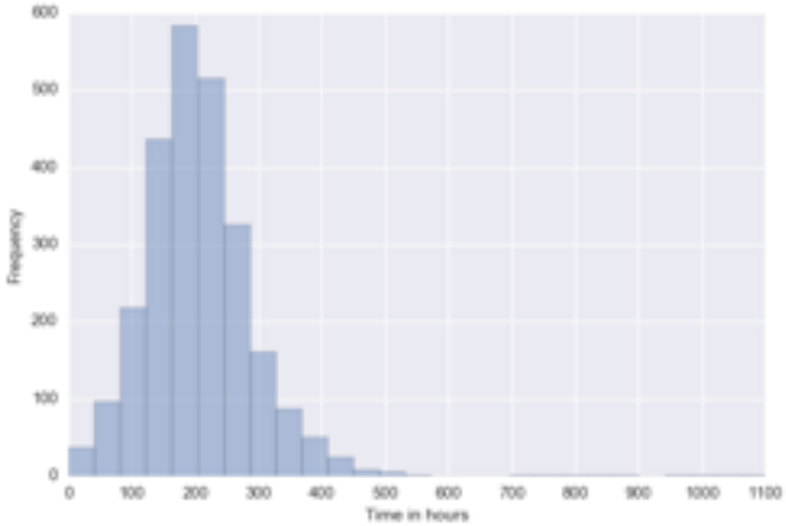
Per determinare gli eventi di RNA editing è necessario disporre sia del genoma che del trascrittoma proveniente dallo stesso individuo, ed un appropriato *workflow* bioinformatico.



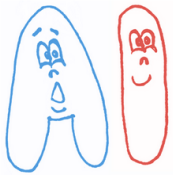
Use Case: RNA editing nei tessuti umani

La predizioni degli eventi di RNA editing richiede molto tempo … perché bisogna esplorare tutte le posizioni genomiche coperte dai trascritti ed applicare diversi filtri per ridurre il tasso di falsi positivi.

```
r1    GGGTGCCTTTATGCAGCAAGGATGCGATATT
r2    GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
r3    GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
r4    GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
r5    GGGTGCCTTTATGCAGCAAGGATGCGATATTTCG
      ..............A......................
      GGGTGCCTTTATGCAGCAAGGATGCGATATTTCGCC gDNA
      ..............G......................
r1    GGGTGCCTTTATGCGGCAAGGATGCGATATT
r2    GGGTGTCTTTATGCAGCAAGGATGCGATACTTCGC
r3    GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
r4    GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
r5    GGGTGCCTTTATGCGGCAAGGATGCGATATTTCG
```



Picardi et al. 2013 Bioinformatics
Picardi et al. 2016 NAR



REDIportal

An ATLAS of A-to-I RNA editing events in human

Problemi più frequenti

- Storage … (80TB per completare il 30% dei dati GTEx)
- Su HTC il numero di ore per processo è limitato
- Su HPC il numero di processi in coda è limitato
- Su HPC non è possibile lanciare un gran numero di processi con richieste esose in termini di RAM e CPU
- Problemi con l'interfaccia di registrazione.

Suggerimenti

- Ampliare le guide per l'utilizzo dei sistemi HPC e HTC
- Aumentare il numero di ore per processo sul sistema HTC
- Maggiori chiarimenti su macchine e risorse disponibili così da poter scegliere la coda più adatta alle proprie esigenze
- Supporto per l'ottimizzazione di tool che richiedono elevato parallelismo
- Sarebbe opportuno avere un unico sistema di gestione delle code
- Miglioramento del sistema di registrazione e login
- Applicare in modo congiunto alle diverse call nazionali o internazionali per rafforzare l'infrastruttura di calcolo ma soprattutto per ampliare lo storage…

Acknowledgments

IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari

PRIN2009
PRIN2010
PRIN2013

Ministero dell'Istruzione dell'Università e Ricerca

CREATIVE-B

SCLEROSI MULTIPLA ONLUS
associazione italiana

MICROMAP
PON Ricerca e Competitività 2007-2013

BioVeL

LifeWatch

CNR-Aging program

ARISLA
http://www.arisla.org/

EPIGEN

elixir

RECAS BARI