

Tutorial per la mappatura di sequenze

*Corso di Laurea Specialistica in Genomica Funzionale, Genomica Computazionale
2025/2026, Fabrizio Mafessoni – tutor Daniela Eugenia Nerelli*

Per domande, lamentele e suggerimenti: fabrizio.mafessoni@units.it

In questa esercitazione vedremo come utilizzare alcuni programmi di mapping. La notazione segue questa convenzione. I comandi sono specificati in Courier New:

```
echo 'this is a command'
```

A volte userò termini come *CARTELLA* e *FILE* come segnaposto per degli esempi (questi non indicano il vero nome di una cartella chiamata *CARTELLA*). Saranno scritti in *MAIUSCOLO CORSIVO*.

```
echo 'UNA_PAROLA'
```

Installazione dei programmi di mapping

Per gli utenti windows e linux

```
sudo apt install bwa samtools bowtie2
```

Per gli utenti Mac

```
brew install bwa samtools bowtie2
```

Se volete usare conda

```
conda install -c bioconda bwa
```

```
conda install -c bioconda samtools
```

```
conda install -c bioconda bowtie2
```

Mapping di reads di *Escherichia coli*

Prima prepariamo la cartella per l'esperimento

```
mkdir ecoli
```

```
cd ecoli
```

```
mkdir fastq
```

```
mkdir reference
```

Scaricate le sequenze fastq di ecoli da mappare

```
wget -P fastq
```

raw.githubusercontent.com/BenLangmead/bowtie/refs/heads/master/reads/e_coli_1000.fq

Notate come usiamo -P con wget per scaricare il file in una specifica cartella (fastq/e_coli_1000.fq)

Scarichiamo il genoma di *Escherichia coli*

```
cd reference

wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000
005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz

gunzip GCF_000005845.2_ASM584v2_genomic.fna.gz

mv GCF_000005845.2_ASM584v2_genomic.fna ecoli_MG1655.fa
```

Ora costruiamo gli indici del genoma con BWA, usando il comando bwa index.

```
bwa index ecoli_MG1655.fa
```

Utilizzando ls potrete notare tutta una serie di file di indici. Come ricorderete dalle lezioni, metodi basati sulla trasformata di Burrow-Wheeler creano due particolari tipi di strutture dati prima della mappatura: il genoma compresso con la trasformata di Burrow-Wheeler e un array di suffissi parziale usato per velocizzare il processo. Potete qui vederli con estensioni .bwt e .sa, rispettivamente!

Ora, possiamo utilizzare questa referencia e questi indici per mappare il nostro dataset di sequenze

Ora potete mappare con bwa-mem. Torniamo alla cartella principale per il progetto ecoli.

```
cd ..
```

La sintassi di bwa mem è semplicemente:

```
bwa mem GENOMADIREFERENZA SEQUENZE_FASTQ
```

Qui stiamo dirigendo l'output su un file sam

```
bwa mem reference/ecoli_MG1655.fa fastq/e_coli_1000.fq >
bwa_ecoli1000.sam
```

Convertiamo il file .sam in .bam con samtools e otteniamo qualche statistica:

```
samtools view -bS bwa_ecoli1000.sam | samtools sort -o
bwa_ecoli1000.bam
```

```
samtools index bwa_ecoli1000.bam
```

```
samtools flagstat bwa_ecoli1000.bam
```

samtools e il formato bam/sam

samtools è di gran lunga uno degli strumenti più usati in bioinformatica, e vi permette di esplorare, filtrare e manipolare files in formato sam e bam. L'opzione view ci permette di estrarre le nostre sequenze, mentre flagstat ci da un sommario e -bS permette di convertire files sam in bam. Inoltre possiamo filtrare delle sequenze sulla base delle flag binarie dei file bam. Ad esempio, per estrarre:

- Mapped reads only

```
samtools view -F 4 bwa_ecoli1000.bam
```

- Unmapped reads only

```
samtools view -f 4 bwa_ecoli1000.bam
```

- Possiamo anche estrarre con l'header incluso con -h o includendo solo l'header:

```
samtools view -f 4 bwa_ecoli1000.bam -H
```

- Reads con ≤ 2 mismatches

Per i curiosi di awk, questo comando cerca il campo NM:i – cioè NM seguito da un intero – e poi ne estrae la fine, che indica con un numero semplicemente il numero dei mismatches (o edit distance):

```
samtools view bwa_ecoli1000.bam | awk '{for(i=12;i<=NF;i++){if($i~/^NM:i:/&& substr($i,6)<=2){print}}}' | head
```

Notate che la colonna NM non è obbligatoria, quindi non necessariamente presente in ogni bam files come le altre.

- Convertire nuovamente un file bam a file sam

```
samtools view -h subset.bam > subset.sam
```

- Vedere dove e quante reads coprono una nostra regione

```
samtools depth -a bwa_ecoli1000.bam | head
```

```
samtools depth -a bwa_ecoli1000.bam | grep -v 0$
```

- Con samtools depth possiamo anche calcolare la coverage media:

```
samtools depth -a bwa_ecoli1000.bam | awk '{sum+=$3; n++} END {print "mean_cov =", sum/n}'
```

- Estrarre la stringa CIGAR

```
samtools view bwa_ecoli1000.bam | cut -f6 | sort | uniq -c | sort -nr
```

Bowtie2

Ora ripetiamo lo stesso con bowtie2. Per bowtie e bowtie2 potrete trovare alcuni indici già creati alle pagine:

Bowtie: <https://bowtie-bio.sourceforge.net/index.shtml>

Bowtie2: <https://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

Da qui potrete trovare gli indirizzi per alcuni genomi con gli indici già creati (così che non dobbiate rigenerarli e possiate risparmiare tempo) e scaricarli. Potete farlo manualmente, ma consiglio di cliccare col tasto destro sulla specie selezionata, andare su “copia link”, e poi scaricare usando wget eseguendo dalla linea di comando:

```
wget INDIRIZZO_DEL_LINK
```

Ad esempio, per gli indici bowtie2 di Arabidopsis thaliana, possiamo eseguire

```
wget https://genome-idx.s3.amazonaws.com/bt/TAIR10.zip
```

Una volta scaricato il file .zip contenente gli indici, potete decomprimerlo con

```
unzip TAIR10.zip
```

Questo creerà una cartella TAIR10 contenente i file .bt2. Noterete anche dei files .rev*.bt2. Questi non sono altro che gli indici per mappare sul reverse complement del genoma! Per alcuni genomi particolarmente grandi gli indici potrebbero avere estensione .bt2l (bt2 large), ma non fa nessuna differenza ai fini dell'utilizzo.

Nel nostro caso, il genoma è piccolo e possiamo costruirli semplicemente da soli.

```
cd reference
```

```
bowtie2-build ecoli_MG1655.fa ecoli_MG1655_bt2
```

```
cd ..
```

```
bowtie2 -x reference/ecoli_MG1655_bt2 -U fastq/  
fastq/e_coli_1000.fq -S bt2_ecoli1000.sam
```

```
samtools view -bS bt2_ecoli1000.sam | samtools sort -o  
bt2_ecoli1000.bam
```

```
samtools index bt2_ecoli1000.bam
```

```
samtools flagstat bt2_ecoli1000.bam
```

Cosa notate, chi mappa di più?

È piuttosto comune che bowtie2 “mappi di più”, in quanto più permissivo. Quindi ci attendiamo più reads mappate a bassa qualità. Verifichiamolo, creando delle tabelle con la mapping quality (MAPQ, campo 5 del formato bam) e il conto delle reads con quella MAPQ

```
samtools view bt2_ecoli1000.bam | awk '{print $5}' | sort |  
uniq -c | awk '{print $2,$1}' | sort -n | column -t  
  
samtools view bwa_ecoli1000.bam | awk '{print $5}' | sort |  
uniq -c | awk '{print $2,$1}' | sort -n | column -t
```

E' importante notare che la MAPQ di bowtie e bwa non sono calcolate allo stesso modo, ed in genere quella di bwa è piu' stringente. Ciò nonostante, nonostante bowtie2 mappi apparentemente piu' reads, possiamo notare come alcuni di queste mappature siano apparentemente di minore qualità.

Gli appassionati di awk (solo se siete interessati ad approfondire awk) avrebbero potuto effettuare tutta l'operazione precedente con il comando awk:

```
samtools view bt2_ecoli1000.bam | awk -v OFS='\t'  
'{h[$5]++}END{for(k in h) print k,h[k]}' | sort -n  
  
samtools view bwa_ecoli1000.bam | awk -v OFS='\t'  
'{h[$5]++}END{for(k in h) print k,h[k]}' | sort -n
```

Esercizi:

1) Mappatura di reads di mais antico con BWA e Bowtie2 sui genomi di sapiens e di mais e compararne la qualità

Scaricate le sequenze da mappare da moodle.

Oppure da github con wget dal link:

```
https://github.com/fabrimafe/GenomicaComputazionale2025_2026/  
raw/refs/heads/main/mais_anticoB_3M.fq.gz
```

Scaricate i genomi di referenza per

Homo sapiens:

```
wget  
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000  
001405.40_GRCh38.p14/GCF_000001405.40_GRCh38.p14_genomic.fna.g  
z
```

Mais

```
wget https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-  
5.0/Zm-B73-REFERENCE-NAM-5.0.fa.gz
```

Ora potete mappare su entrambi seguendo le istruzioni viste prima

- 2) Prima di mappare, i genomi vengono spesso mascherati (le zone ripetitive sono mascherate con delle N). Ad esempio, possiamo esplorare i genomi di Arabidopsis**

[Index of /pub/plants/release-62/fasta/arabidopsis_thaliana/dna](https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-62/fasta/arabidopsis_thaliana/dna)

Non mascherato

```
wget https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-62/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz
```

Mascherato soft (soft masking)

```
wget https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-62/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna_sm.toplevel.fa.gz
```

Mascherato pienamente (hard masking)

```
wget https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-62/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna_rm.toplevel.fa.gz
```

Mascherate le sequenze ripetute (microsatelliti binucleotidici semplici come TATATA..) usando i comandi linux di vostra conoscenza il genoma di mais, e mappate sulla versione mascherata.

- 3) Mappate le reads di E.coli sui genomi batterici (individualmente, possibilmente usando dei loops) di:**

Salmonella Enterica

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/945/GCF_000006945.2_ASM694v2/GCF_000006945.2_ASM694v2_genomic.fasta.gz
```

Pseudomonas aeruginosa

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/765/GCF_000006765.1_ASM676v1/GCF_000006765.1_ASM676v1_genomic.fasta.gz
```

Shigella flexneri

```
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/925/GCF_000006925.2_ASM692v2/GCF_000006925.2_ASM692v2_genomic.fasta.gz
```

- Dove mappano meglio queste reads?
- Provate successivamente (dopo aver mappato individualmente) a creare una referenza concatenata usando i comandi linux di vostra conoscenza. Indicizzate questa referenza concatenata ed effettuate competitive mapping. Cosa notate?

4) **Esistono vari strains di E.coli. Da quale ceppo venivano le reads che abbiamo precedentemente usato. Mappate prima le reads di E.coli sui vari genomi di E.coli individualmente, possibilmente usando dei loops:**

- MG1655 (standard)

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz
gunzip GCF_000005845.2_ASM584v2_genomic.fna.gz
mv GCF_000005845.2_ASM584v2_genomic.fna ecoli_MG1655.fa
```

- O157:H7 Sakai

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/008/865/GCF_000008865.2_ASM886v2/GCF_000008865.2_ASM886v2_genomic.fna.gz
gunzip GCF_000008865.2_ASM886v2_genomic.fna.gz
mv GCF_000008865.2_ASM886v2_genomic.fna ecoli_O157H7.fa
```

- BL21(DE3)

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/565/GCF_000009565.1_ASM956v1/GCF_000009565.1_ASM956v1_genomic.fna.gz
gunzip GCF_000009565.1_ASM956v1_genomic.fna.gz
mv GCF_000009565.1_ASM956v1_genomic.fna ecoli_BL21.fa
```

- W3110: patogenico

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/010/245/GCF_000010245.2_ASM1024v1/GCF_000010245.2_ASM1024v1_genomic.fna.gz
gunzip GCF_000010245.2_ASM1024v1_genomic.fna.gz
mv GCF_000010245.2_ASM1024v1_genomic.fna ecoli_W3110.fa
```

- UPEC CFT073: patogenico

wget

```
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/007/445/GCF_000007445.1_ASM744v1/GCF_000007445.1_ASM744v1_genomic.fna.gz
```

```
gunzip GCF_000007445.1_ASM744v1_genomic.fna.gz
```

```
mv GCF_000007445.1_ASM744v1_genomic.fna ecoli_CFT073.fa
```

Cosa notate?

- Provate successivamente (dopo aver mappato individualmente) a creare una referenza concatenata usando i comandi linux di vostra conoscenza. Indicizzate questa referenza concatenata ed effettuate competitive mapping. A quale ceppo appartenevano le nostre reads?