

# Algoritmi per la trasformata di Burrows-Wheeler posizionale con compressione run-length

Davide Cozzi

**Relatore:** *Prof. Raffaella Rizzi*    **Correlatore:** *Dr. Yuri Pirola*

*Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)  
Università degli Studi di Milano Bicocca*

25 Ottobre 2022

# Outline

- 1 Introduzione
- 2 Preliminari
- 3 Metodo
- 4 Risultati sperimentali
- 5 Conclusioni e sviluppi futuri

# Un punto di vista per il pangenoma

Negli ultimi anni si è assistito a un cambio di paradigma nel campo della *bioinformatica*, ovvero il passaggio dallo studio della sequenza lineare di un singolo genoma a quello di un insieme di genomi, provenienti da un gran numero di individui, al fine di poter considerare anche le varianti geniche. Questo nuovo concetto è stato introdotto da Tettelin, nel 2005, con il termine di **pangenoma**.

Uno degli approcci più usati per rappresentare il **pangenoma** è attraverso un pannello di aplotipi, ovvero, da un punto di vista computazionale, una matrice di M righe, corrispondenti agli individui, e N colonne, corrispondenti ai siti con le varianti.

Un **aplotipo** è l'insieme di alleli, ovvero di varianti che, a meno di mutazioni, un organismo eredita da ogni genitore.

# RLBWT

## Esempio

| Thresholds |   |    |           |     |           |                   |                   |
|------------|---|----|-----------|-----|-----------|-------------------|-------------------|
| A          | T | SA | SA sample | BWT | Run heads | LCP               | $\mathcal{M}$     |
| *          | * | 15 | 15        | A   | A         | 0                 | \$ATTAGATTACATTA  |
|            |   | 14 | 14        | T   | T         | 0                 | A\$ATTAGATTACATT  |
|            |   | 9  |           | T   |           | 1                 | ACATTAS\$ATTAGATT |
|            |   | 4  |           | T   |           | 1                 | AGATTACATTAS\$ATT |
|            |   | 11 | 11        | C   | C         | 1                 | ATTA\$ATTAGATTAC  |
|            |   | 6  | 6         | G   | G         | 4                 | ATTACATTAS\$ATTAG |
|            |   | 1  | 1         | \$  | \$        | 4                 | ATTAGATTACATTAS\$ |
| 10         |   | 10 | A         | A   | 0         | CATTAS\$ATTAGATTA |                   |
| 5          |   |    | A         |     | 0         | GATTACATTAS\$ATTA |                   |
| 13         |   | 13 | T         | T   | 0         | TA\$ATTAGATTACAT  |                   |
| 8          |   |    | T         |     | 2         | TACATTAS\$ATTAGAT |                   |
| 3          |   |    | T         |     | 2         | TAGATTACATTAS\$AT |                   |
| 12         |   | 12 | A         | A   | 1         | TTAS\$ATTAGATTACA |                   |
| 7          |   |    | A         |     | 3         | TTACATTAS\$ATTAGA |                   |
| 2          |   |    | A         |     | 3         | TTAGATTACATTAS\$A |                   |

# MONI e PHONI

## MONI

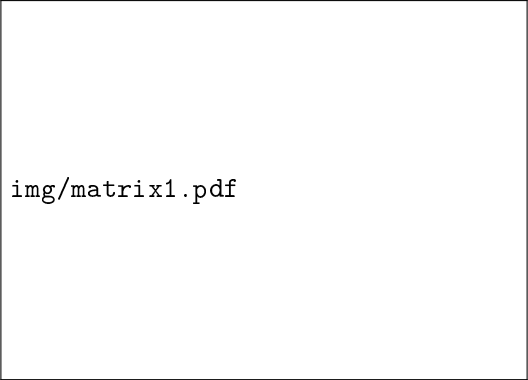
Rossi et al., nel 2021, sfruttarono le conoscenze relative alla RLBWT e all'r-index per ideare **MONI**. In questa soluzione si ha la costruzione, in due sweep, tramite l'uso delle threshold (*algoritmo di Bannai*), dell'array delle matching statistics, da cui si computano i maximal exact match.

**Rossi et al:** *MONI: A pangenomic index for finding maximal exact matches, 2021*

## PHONI

Nel 2021, Boucher, Gagie, Rossi et al. proposero un ulteriore miglioramento di quanto fatto in MONI, con **PHONI**, usando le LCE query al posto delle threshold, ottenendo un algoritmo “online”.

**Boucher et al:** *PHONI: Streamed matching statistics with multi-genome references, 2021*



img/matrix1.pdf

$a_6 = [14, 15, 0, 9, 10, 16, 8, 11, 12, 13, 18, 19, 1, 2, 3, 17, 4, 5, 6, 7]$

$d_6 = [6, 0, 4, 2, 0, 0, 5, 0, 0, 0, 3, 0, 4, 0, 0, 6, 4, 0, 0, 0]$

**Durbin:** *Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT)*, 2014

# Set-maximal exact match

Calcolo SMEM via **algoritmo 5 di Durbin** in tempo  $\mathcal{O}(NM) + \text{Avg.}\mathcal{O}(N + c)$ ,  
richiedendo  $13NM$  byte

`img/pbwtmatch.pdf`

# Componenti e strutture dati, una panoramica

img/ds.pdf



`img/make_time_mem_paper.png`

## Performance calcolo degli SMEM con 100 query

`img/exe_time_mem_paper.png`

## Performance calcolo degli SMEM per singole query

`img/exe_time_single_paper.png`

# Considerazioni e sviluppi

## Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obiettivo della tesi, ovvero lo sviluppo di un algoritmo, efficiente in spazio, per il calcolo degli SMEM di un aplotipo esterno contro un pannello, è stato raggiunto con risultati molto interessanti

# Considerazioni e sviluppi

## Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obiettivo della tesi, ovvero lo sviluppo di un algoritmo, efficiente in spazio, per il calcolo degli SMEM di un aplotipo esterno contro un pannello, è stato raggiunto con risultati molto interessanti

## Sviluppi futuri

- SMEM interni con RLPBWT
- RLPBWT multiallelica
- RLPBWT per pannelli con dati mancanti
- calcolo K-SMEM con RLPBWT

# Considerazioni e sviluppi

## Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obiettivo della tesi, ovvero lo sviluppo di un algoritmo, efficiente in spazio, per il calcolo degli SMEM di un aplotipo esterno contro un pannello, è stato raggiunto con risultati molto interessanti

## Sviluppi futuri

- SMEM interni con RLPBWT
- RLPBWT multiallelica
- RLPBWT per pannelli con dati mancanti
- calcolo K-SMEM con RLPBWT

**Bonizzoni, Boucher, Cozzi, Gagie, Kashgouli, Köppl e Rossi:**

*Compressed data structures for population-scale positional Burrows–Wheeler transforms, 2022*

Grazie per l'attenzione

Davide Cozzi

**Relatore:** *Prof. Raffaella Rizzi*    **Correlatore:** *Dr. Yuri Pirola*

*Dipartimento di Informatica, Sistemistica e Comunicazione (DISCo)  
Università degli Studi di Milano Bicocca*

25 Ottobre 2022