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

Davide Cozzi

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

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

26 Ottobre 2022

Outline

Il pangenoma

Un cambio di paradigma

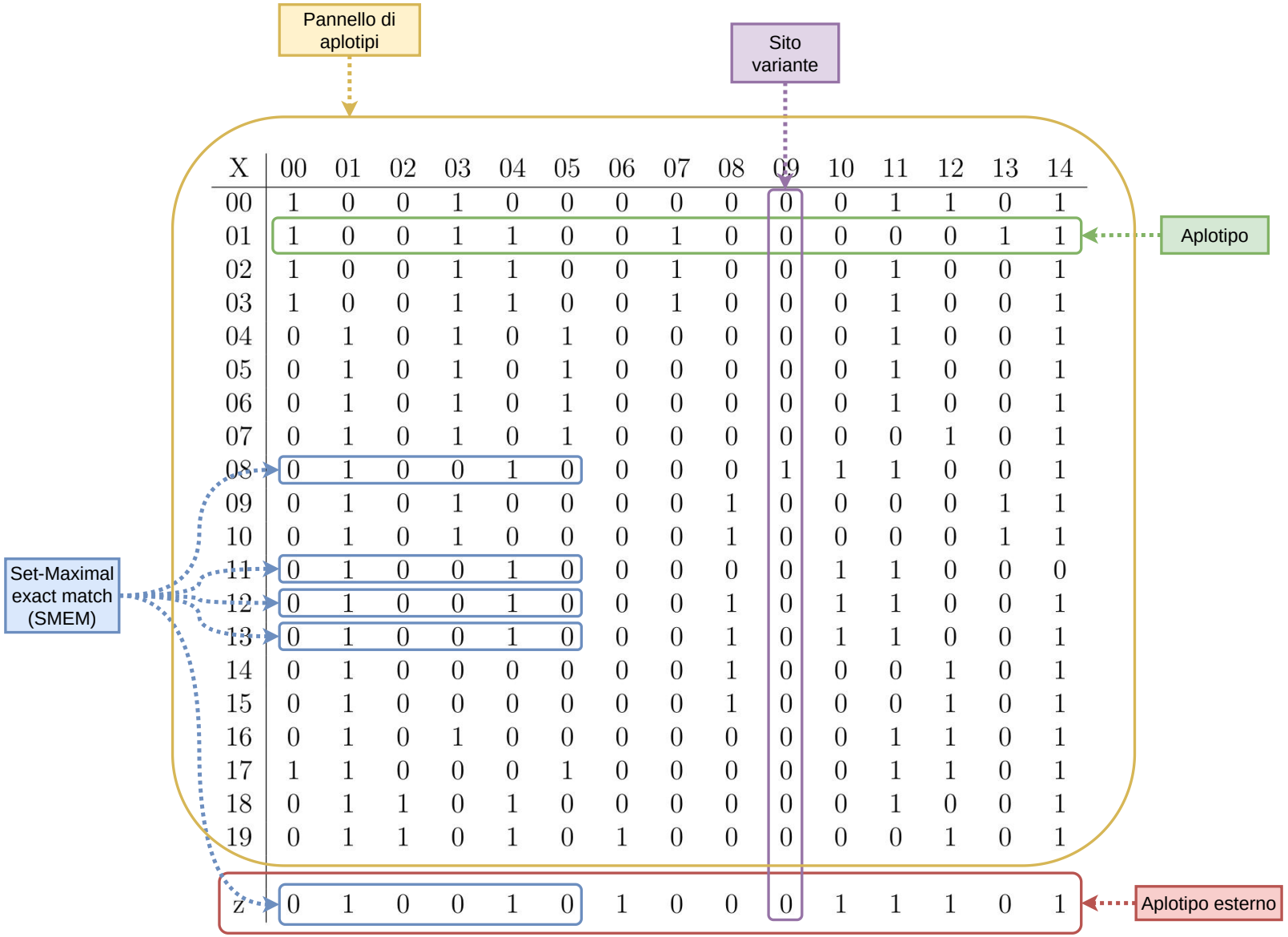
- singolo genoma $\Longrightarrow$ insieme di genomi
- studio delle varianti geniche tra genomi di diversi individui

Rappresentazioni del pangenoma

- grafo del pangenoma
- **pannello di aplotipi**

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

Il pangenoma



X	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1
01	1	0	0	1	1	0	0	1	0	0	0	0	0	1	1
02	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
03	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1
08	0	1	0	0	1	0	0	0	0	1	1	1	0	0	1
09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
11	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
12	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
13	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1
17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1
18	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1
19	0	1	1	0	1	0	1	0	0	0	0	0	1	0	1
z	0	1	0	0	1	0	1	0	0	0	1	1	1	0	1

Trasformata di Burrows–Wheeler posizionale

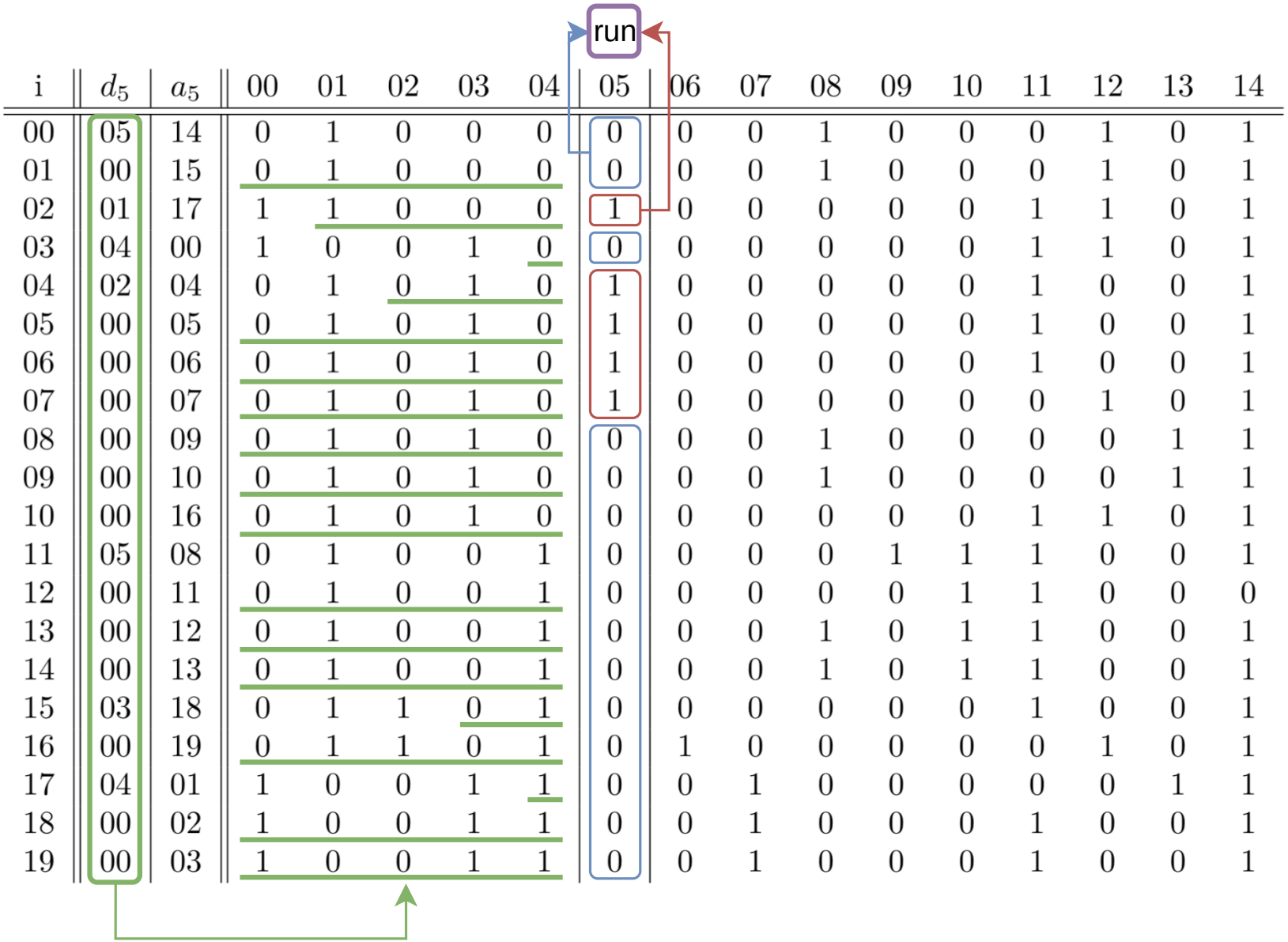
PBWT - *Durbin, Bioinformatics, 2014*

Dato pannello di M aplotipi, lunghi N siti (biallelici: $\Sigma = \{0, 1\}$), si definisce PBWT del pannello una collezione di $N + 1$ coppie di array (a_k, d_k), $0 \leq k \leq N$, dove:

- a_k è il **prefix array** della colonna k
- d_k è il **divergence array** della colonna k

- la PBWT è basata sul riordinamento co-lessicografico a ogni colonna
- il pannello, riordinato in ogni colonna k con a_k, è detto matrice PBWT
- aplotipi simili, riordinati consecutivamente alla colonna k, è molto probabile presentino il medesimo carattere in colonna $k + 1$

Trasformata di Burrows–Wheeler posizionale



i	d_5	a_5	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
00	05	14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
01	00	15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
02	01	17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1
03	04	00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1
04	02	04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
05	00	05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
06	00	06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
07	00	07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1
08	00	09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
09	00	10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
10	00	16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1
11	05	08	0	1	0	0	1	0	0	0	0	1	1	1	0	0	1
12	00	11	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
13	00	12	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
14	00	13	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
15	03	18	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1
16	00	19	0	1	1	0	1	0	1	0	0	0	0	0	1	0	1
17	04	01	1	0	0	1	1	0	0	1	0	0	0	0	0	1	1
18	00	02	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
19	00	03	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1

Scopo della tesi

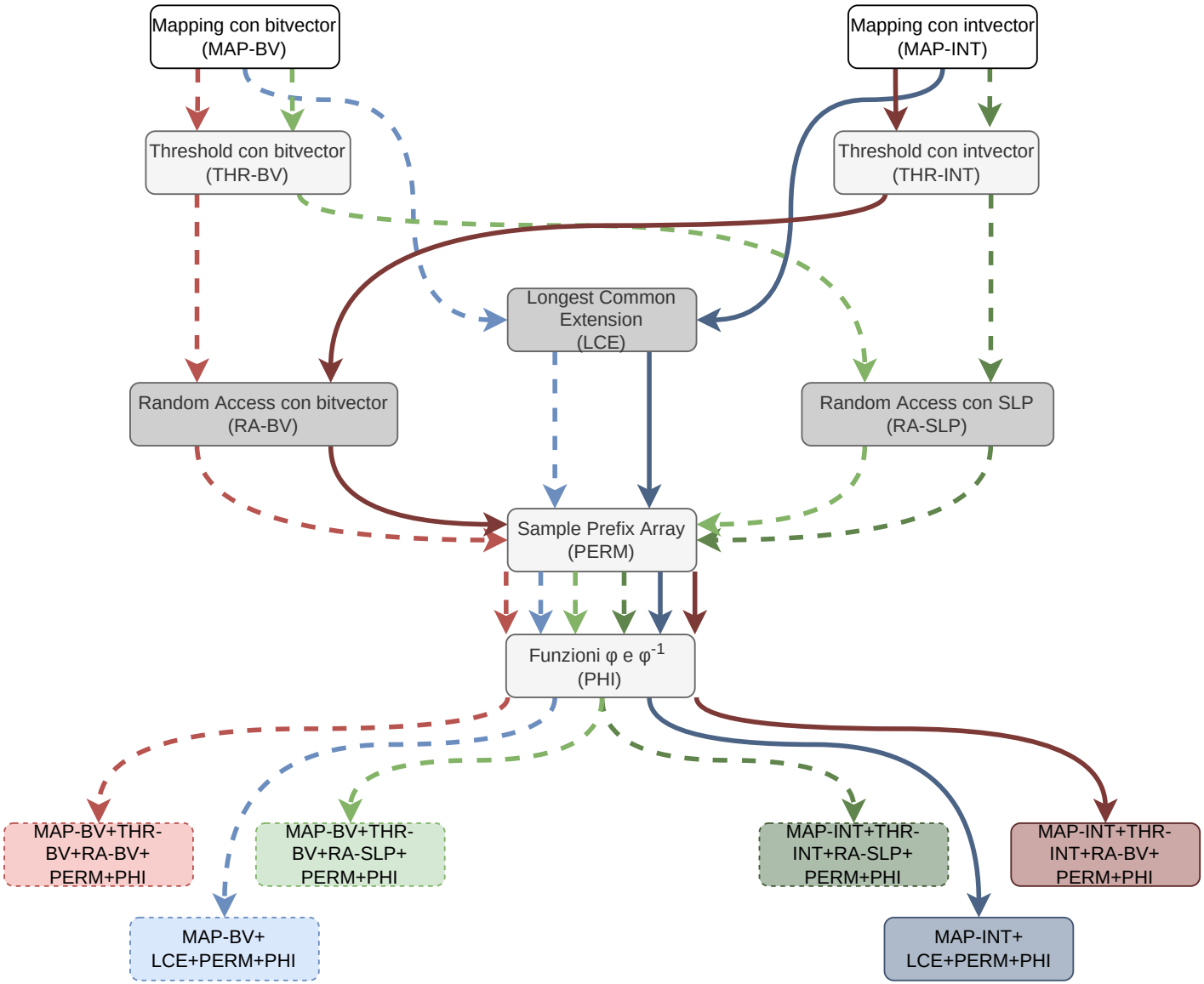
Lo scopo di questa tesi è stato quello di creare una variante run-length encoded della PBWT (RLPBWT) che permettesse, in modo efficiente dal punto di vista della memoria richiesta, il calcolo degli SMEM con aplotipo esterno.

Complessità calcolo degli SMEM con un algoritmo naïve: $\mathcal{O}(N^2M)$

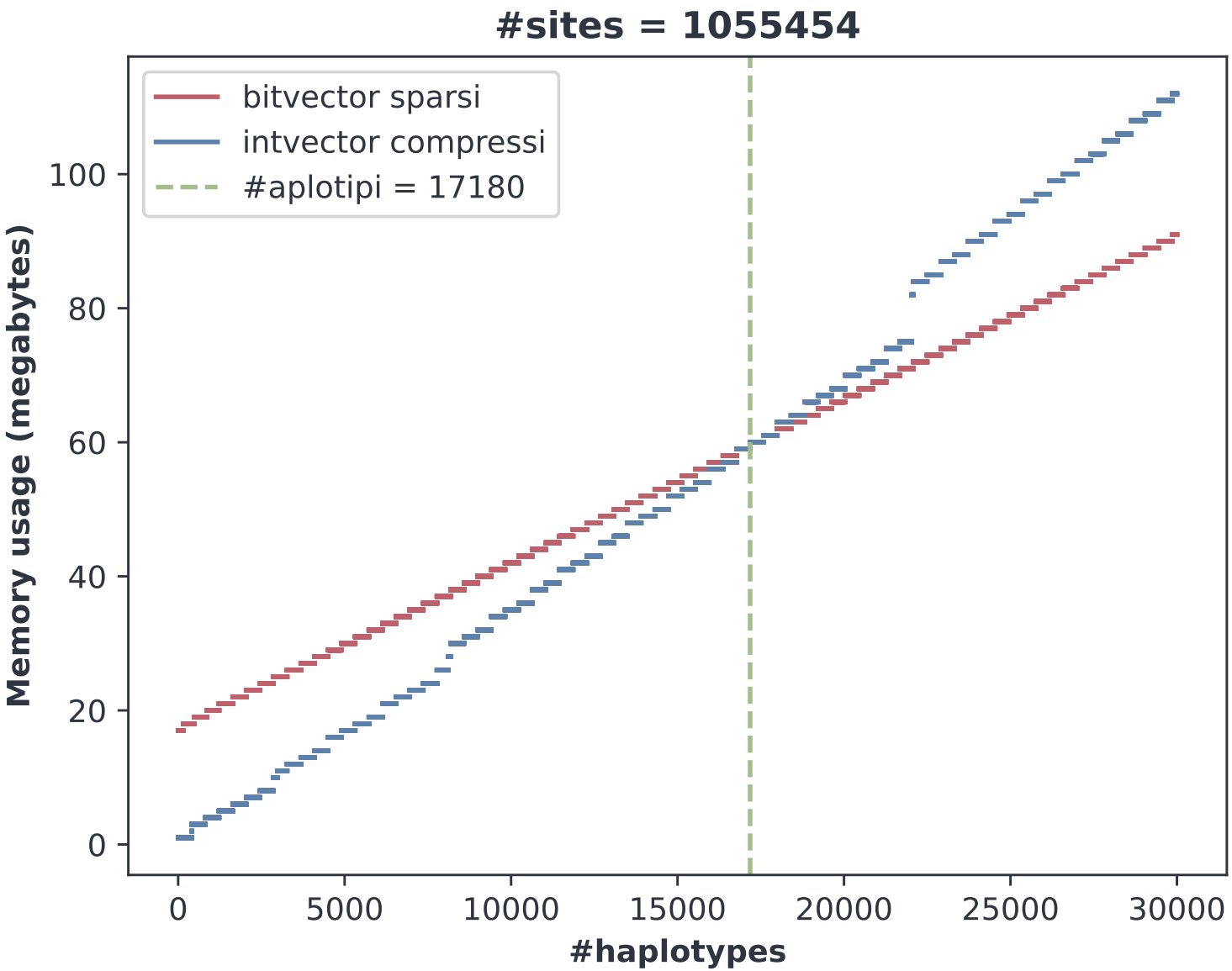
Calcolo degli SMEM con aplotipo esterno per Durbin:

- tempo: $\mathcal{O}(NM)$ + Avg. $\mathcal{O}(N + c)$
- spazio: $\mathcal{O}(NM) \implies \underline{13NM \textbf{ byte}}$

Sottostrutture e strutture dati composte per la RLPBWT



Qualche confronto in spazio



Calcolo degli SMEM

Matching statistics per la PBWT

Dato un pannello $X = \{x_0, \ldots, x_{M-1}\}$, $x_i = N$, e un aplotipo esterno z, tale che $|z| = N$, si definisce **matching statistics** di z su X un array MS di coppie (row, len), $|\text{MS}| = N$, tale che:

- $x_{\text{MS}[i].\text{row}}[i - \text{MS}[i].\text{len} + 1, i] = z[i - \text{MS}[i].\text{len} + 1, i]$, ovvero si ha che l'aplotipo esterno ha un match, lungo MS$[i]$.len, terminante in colonna i, con la riga MS$[i]$.row-esima del pannello
- $z[i - \text{MS}[i].\text{len}, i]$ non è un suffisso terminante in colonna i di un qualsiasi sottoinsieme di righe di X

Calcolo degli SMEM

Matching statistics per la PBWT

Dato un pannello $X = \{x_0, \ldots, x_{M-1}\}$, $x_i = N$, e un aplotipo esterno z, tale che $|z| = N$, si definisce **matching statistics** di z su X un array MS di coppie (row, len), $|\text{MS}| = N$, tale che:

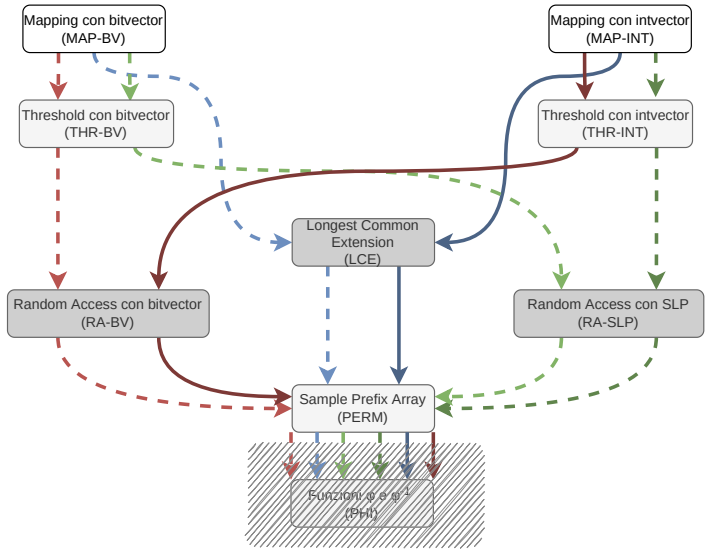
- $x_{\text{MS}[i].\text{row}}[i - \text{MS}[i].\text{len} + 1, i] = z[i - \text{MS}[i].\text{len} + 1, i]$, ovvero si ha che l'aplotipo esterno ha un match, lungo MS$[i]$.len, terminante in colonna i, con la riga MS$[i]$.row-esima del pannello
- $z[i - \text{MS}[i].\text{len}, i]$ non è un suffisso terminante in colonna i di un qualsiasi sottoinsieme di righe di X

SMEM da MS

Dato un array di matching statistics MS si ha che $z[i - l + 1, i]$ presenta uno SMEM di lunghezza l con la riga MS$[i]$.row-esima del pannello X sse: $\text{MS}[i].\text{len} = l \wedge (i = N - 1 \vee \text{MS}[i].\text{len} \geq \text{MS}[i + 1].\text{len})$

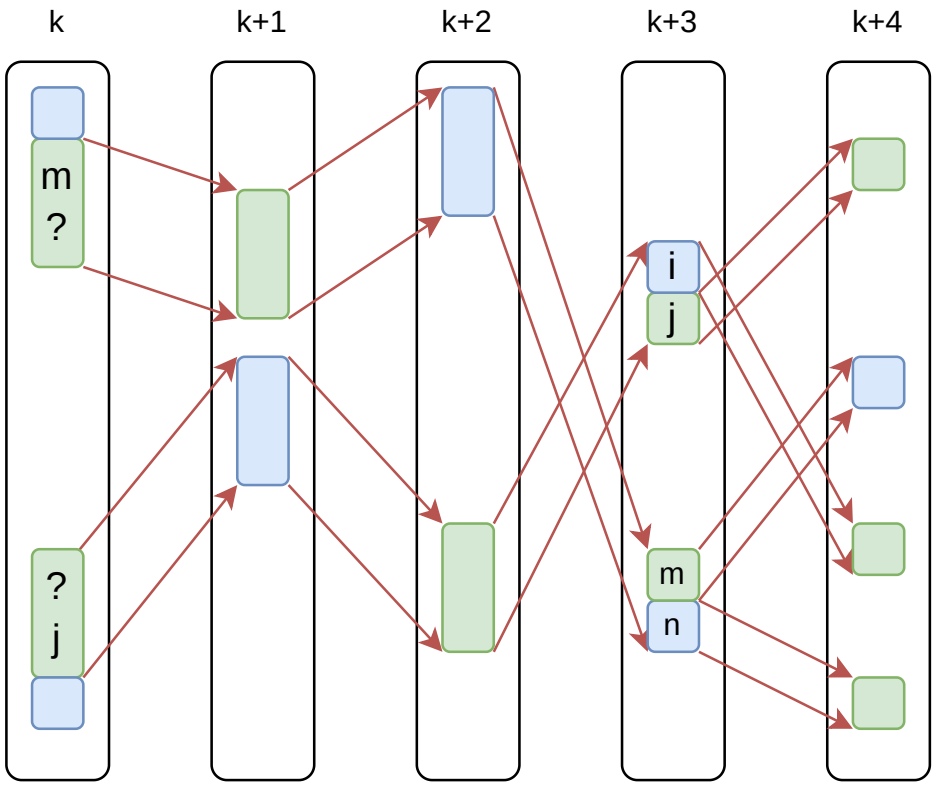
Esempio di matching statistics

X	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
00	1	0	0	1	0	0	0	0	0	0	0	1	1	0	1
01	1	0	0	1	1	0	0	1	0	0	0	0	0	1	1
02	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
03	1	0	0	1	1	0	0	1	0	0	0	1	0	0	1
04	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
05	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
06	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1
07	0	1	0	1	0	1	0	0	0	0	0	0	1	0	1
08	0	1	0	0	1	0	0	0	0	1	1	1	0	0	1
09	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
10	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1
11	0	1	0	0	1	0	0	0	0	0	1	1	0	0	0
12	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
13	0	1	0	0	1	0	0	0	1	0	1	1	0	0	1
14	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
15	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1
16	0	1	0	1	0	0	0	0	0	0	0	1	1	0	1
17	1	1	0	0	0	1	0	0	0	0	0	1	1	0	1
18	0	1	1	0	1	0	0	0	0	0	0	1	0	0	1
19	0	1	1	0	1	0	1	0	0	0	0	0	1	0	1
z	0	1	0	0	1	0	1	0	0	0	1	1	1	0	1

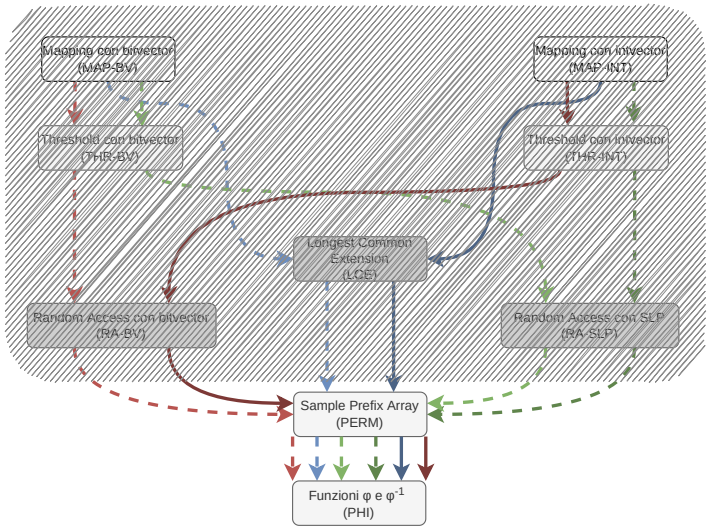


k	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14
row	19	19	16	15	13	13	19	19	19	19	11	11	17	17	17
len	1	2	3	4	5	6	4	5	6	7	4	5	2	3	4

Struttura per le funzioni φ e φ^{-1}



...



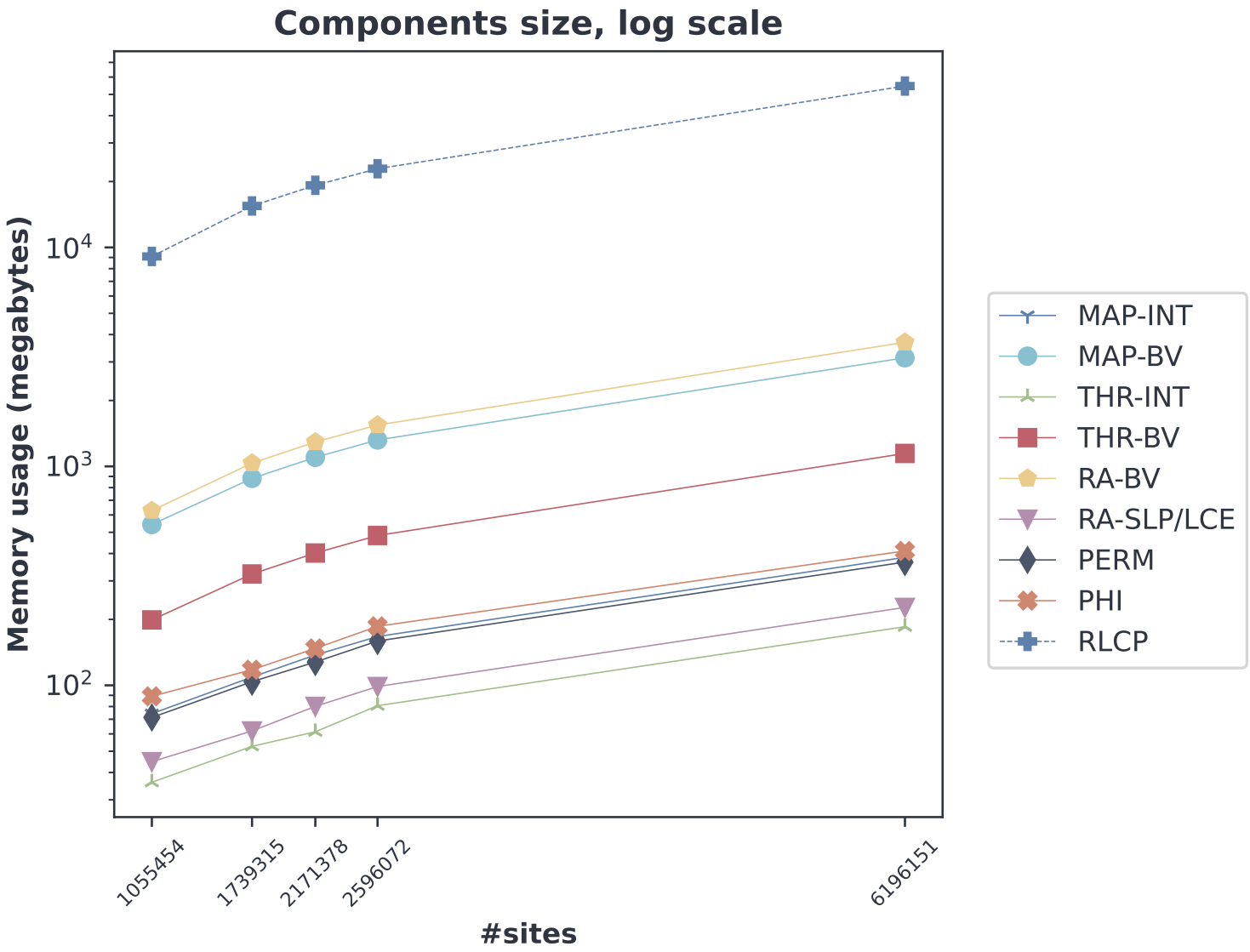
Dati di input

Pannelli di varianti reali

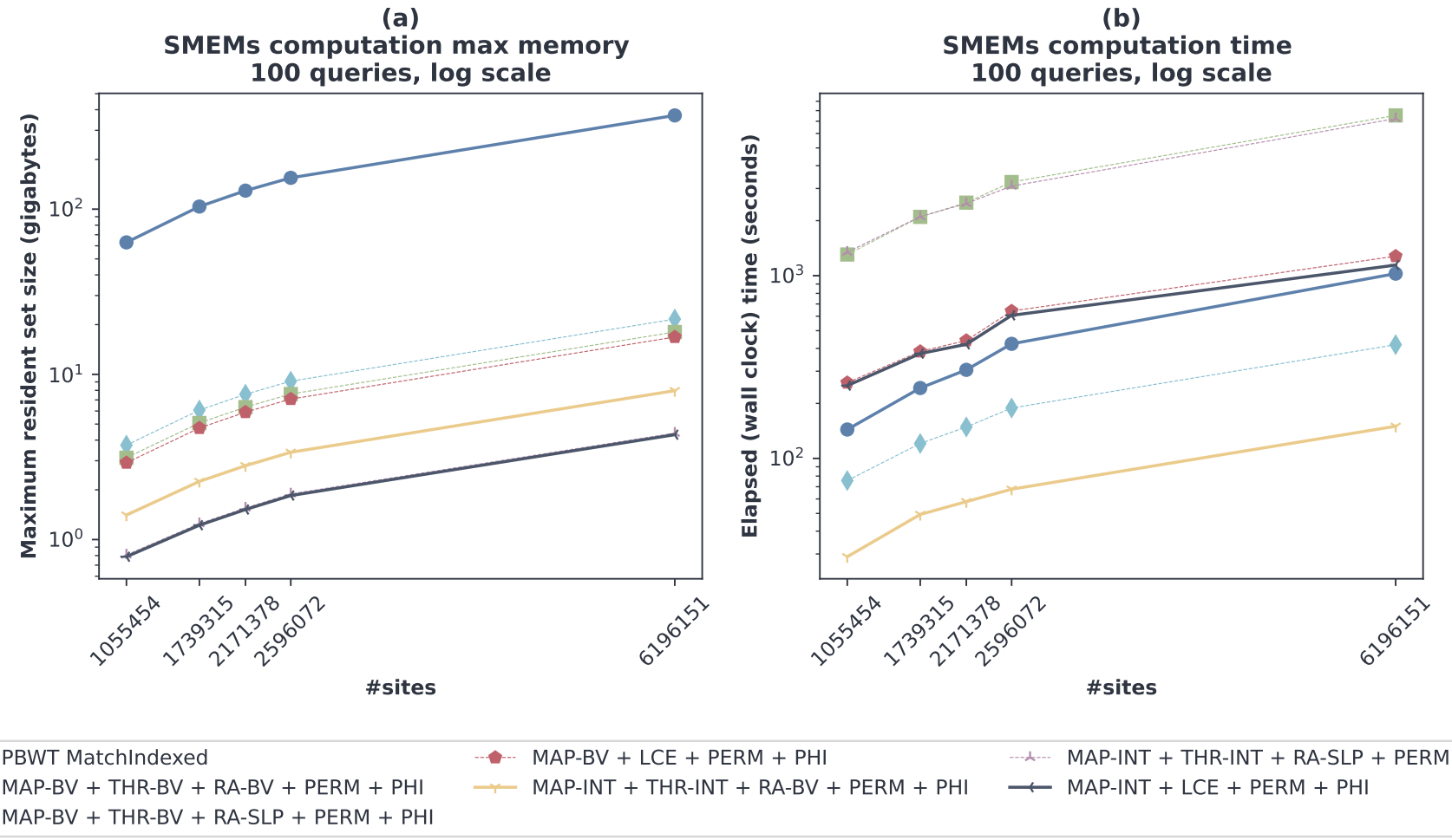
- dati reali relativi alla phase 3 del **1000 Genome Project**
- numero costante di aplotipi: 5008
- estrazione di 100 query $\Longrightarrow$ pannelli con 4908 aplotipi
- numero variabile di siti

Chr	#Siti	Media run
chr22	1.055.454	14
chr20	1.739.315	11
chr18	2.171.378	11
chr16	2.596.072	12
chr1	6.196.151	11

Costo in memoria delle componenti



Performance calcolo degli SMEM con 100 query



Considerazioni e sviluppi futuri

Alcune considerazioni

- le strutture dati e gli algoritmi proposti hanno confermato la potenzialità dell'uso di strutture run-length encoded in pangenomica
- l'obbiettivo 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

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

Ulteriori dettagli

Bonizzoni, Boucher, Cozzi, Gagie, Kashgouli, Köppl e Rossi:
Compressed data structures for population-scale positional Burrows–Wheeler transforms, *bioRxiv, 2022*

17th Workshop on Compression, Text and Algorithms (WCTA), 11 Novembre 2022

Grazie per l'attenzione

Mapping e threshold

Colonna matrice PBWT

$$y^5 = 00101111000000000000$$

intvector compressi

$$p_5 = [0, 2, 3, 4, 8]$$

$$uv_5 = [0, 2, 1, 3, 5], \quad c[5] = 15, \quad start_5 = \top$$

$$t_5 = [0, 3, 3, 4, 11]$$

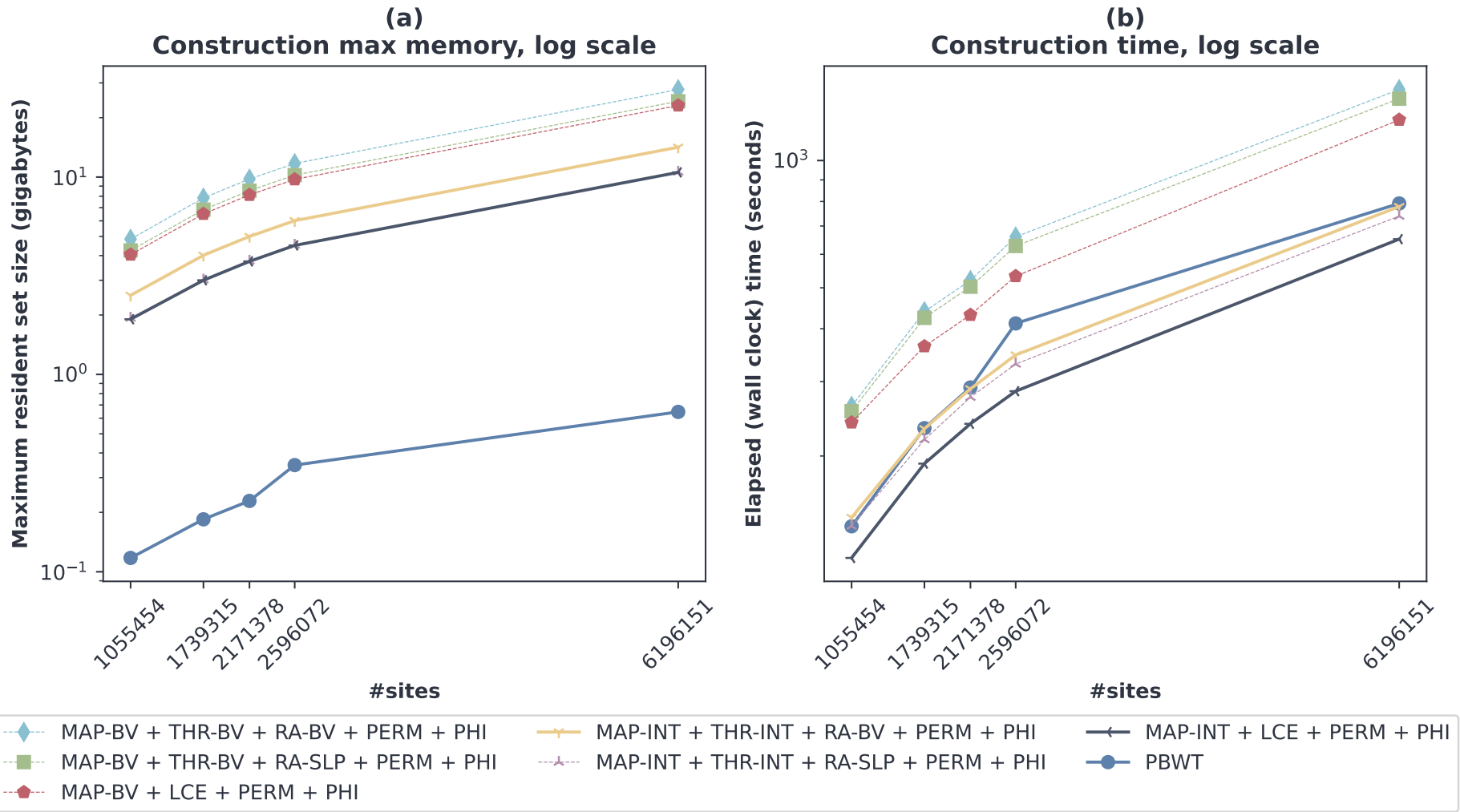
bitvector sparsi

$$h_5 = 01110001000000000001$$

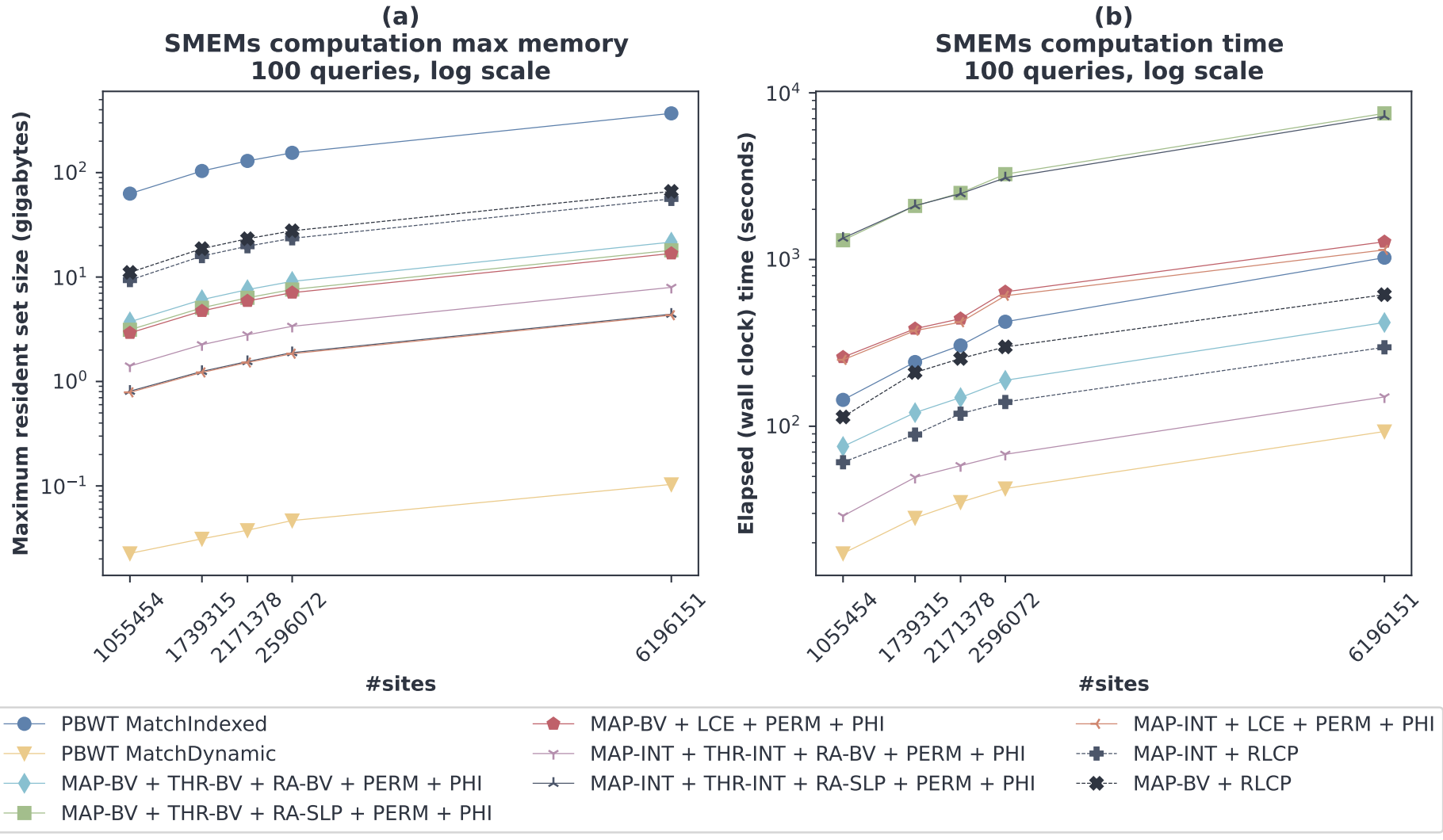
$$u_5 = 011000000000001, \quad v_5 = 10001, \quad c[5] = 15, \quad start_5 = \top$$

$$t_5 = 10111000000100000000$$

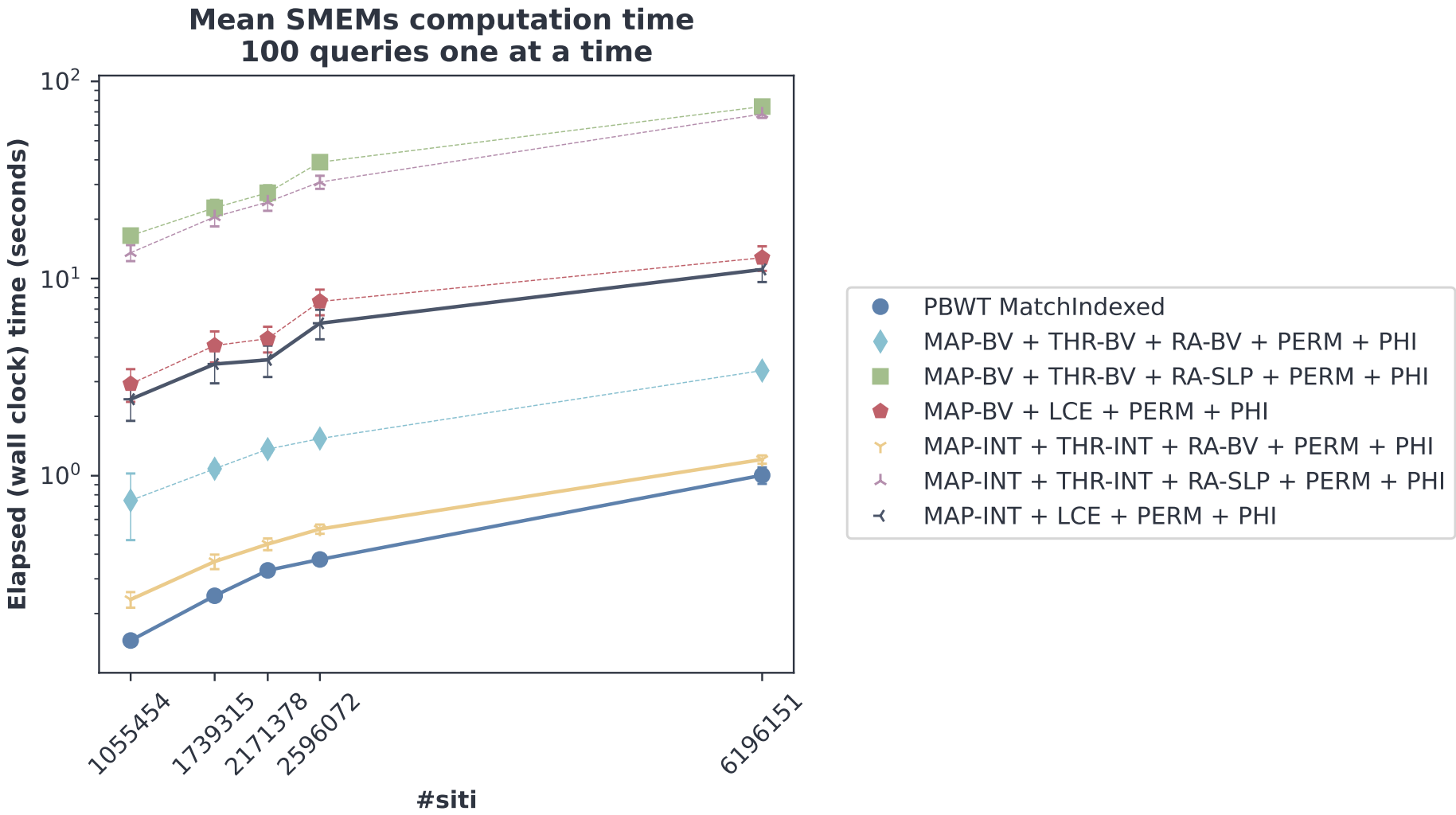
Performance costruzione strutture dati



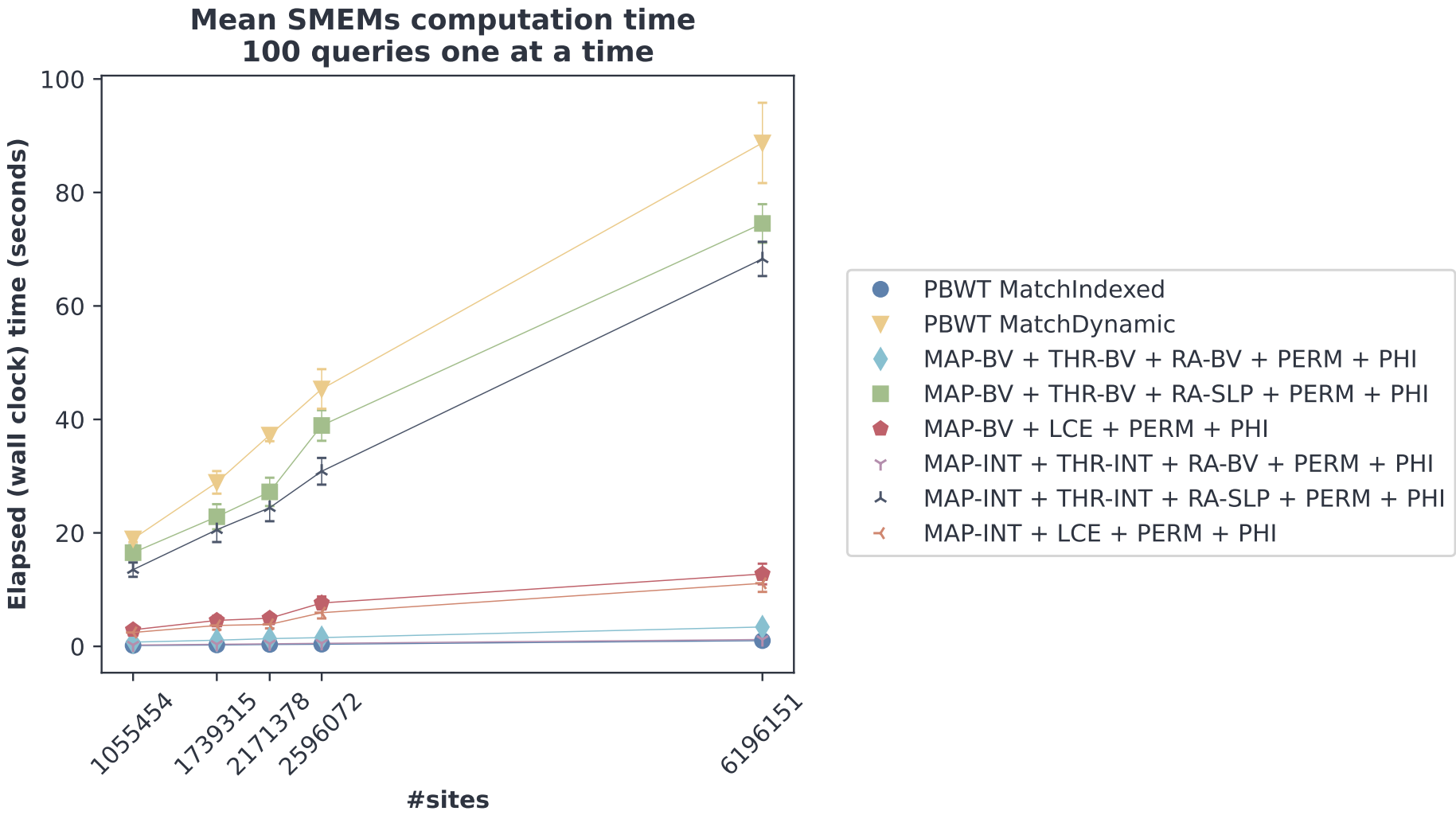
Performance calcolo degli SMEM con 100 query



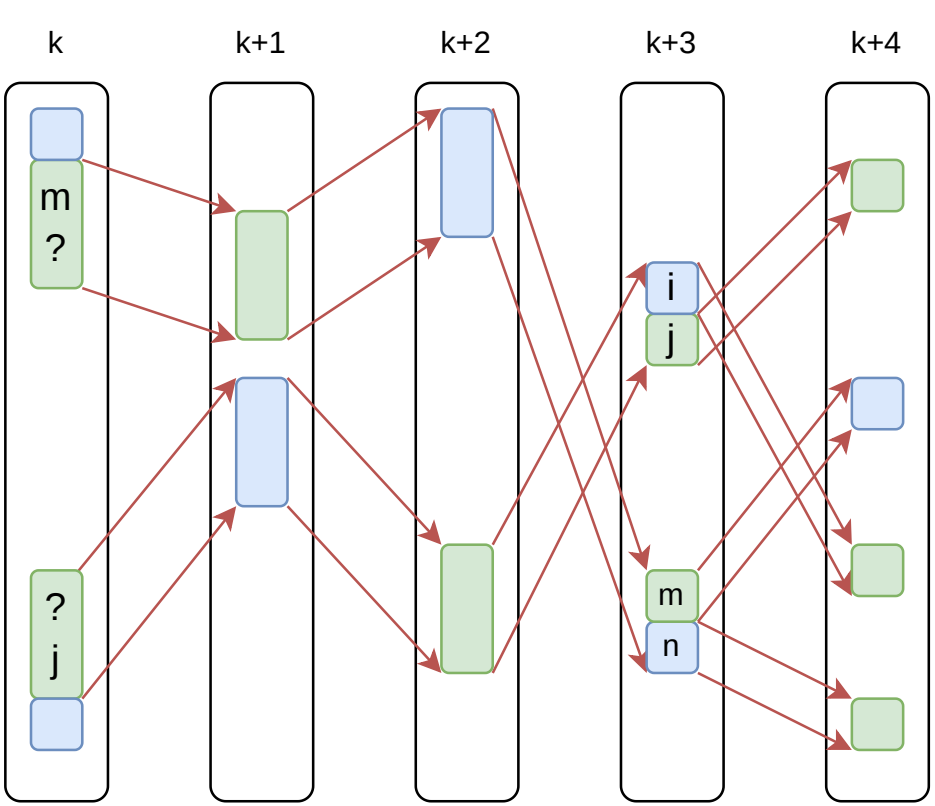
Performance calcolo degli SMEM per singole query



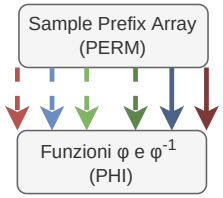
Performance calcolo degli SMEM per singole query



Struttura per le funzioni φ e φ^{-1}



...



$$\Phi_j = [0, 0, 0, 1, 0, \ldots], \quad \Phi_m^{-1} = [0, 0, 0, 1, 0, \ldots], \quad \Phi_{supp} = [i, \ldots], \quad \Phi_{supp}^{-1} = [n, \ldots]$$

$$\Phi^j_{supp}[\text{rank}^{\varphi}_j(0)] = \Phi^j_{supp}[0] = i, \quad \Phi^{-1\ m}_{supp}[\text{rank}^{\varphi^{-1}}_m(0)] = \Phi^{-1\ m}_{supp}[0] = n$$