# Supplementary Material for "Haplotype-aware sequence alignment to pangenome graphs"

## Note S1: Algorithm for solving Problem 5 (Gap-sensitive co-linear chaining)

To efficiently compute gap costs between anchors during co-linear chaining, we follow the idea as described in (Chandra and Jain, 2023). We assume the same definition of gap cost function as used in the Problem 2a of reference (Chandra and Jain, 2023). First, we precompute an index of the DAG to facilitate efficient calculation of gaps during chaining. Let $D(v_1, v_2)$ denote the shortest path measured in terms of the number of characters from $v_1$ to $v_2$. If path $P_h$ covers vertex $v$, then suppose $dist2begin(v, h)$ is the number of characters in haplotype $h$ before vertex $v$. Accordingly, our precomputed index contains $last2reach(v, h)$, $D(last2reach(v, h), v)$ and $dist2begin(v, h)$ for all $v \in V$ and $h \in [|\mathcal{H}|]$. All these quantities can be computed in $O(|E||\mathcal{H}|)$ time during preprocessing of the DAG (Chandra and Jain, 2023). See Algorithm 2 for an outline of the gap-sensitive haplotype-aware chaining procedure. This is similar to Algorithm 1 except that we introduce two variables $g_1$ and $g_2$ (Lines 18, 27) to consider gaps between anchors. The arguments in the proofs of Lemma 6 in this paper and Lemma 4 in reference (Chandra and Jain, 2023) can be easily extended to argue the correctness of Algorithm 2.

**Algorithm 2:** $O(|\mathcal{H}|N \log |\mathcal{H}|N)$ time chaining algorithm for Problem 5

**Input:** Weighted anchors $M[1..N]$, haplotype paths $P_1, \ldots, P_{|\mathcal{H}|}$, precomputed index of DAG, parameter $\gamma$
**Output:** Table $C$ such that $C(i,h)$ = optimal score of a chain that ends at $(i,h)$, for all $i \in [N], h \in [|\mathcal{H}|]$

Initialize search trees $\mathcal{T}_h$, for all $h \in [|\mathcal{H}|]$, using keys $\{M[i].d \mid 1 \leq i \leq N\}$ and values $-\infty$
Initialize $C(i,h), C''(i)$ and $C_{l2r}(i)$ as $weight(M[i])$, for all $i \in [N], h \in [|\mathcal{H}|]$
/* Create array $Z$ that stores tuples of the form $(v, pos, task, anchor, haplotype)$, where $v \in V$, $pos \in \mathbb{N}$, $task \in \{0,1,2\}$, $anchor \in [N]$ and $haplotype \in [|\mathcal{H}|]$.*/
**for** $i \leftarrow 1$ ***to*** $N$ **do**
  **for** $h \leftarrow 1$ ***to*** $|\mathcal{H}|$ **do**
    **if** $h \in haps(M[i].v)$ **then**
      $Z.push(M[i].v, M[i].x, 0, i, h)$
      $Z.push(M[i].v, M[i].x, 1, i, h)$
      $Z.push(M[i].v, M[i].y, 2, i, h)$
    **else if** $last2reach(M[i].v, h)$ *exists* **then**
      $v_{l2r} \leftarrow last2reach(M[i].v, h)$
      $Z.push(v_{l2r}, |\sigma(v_{l2r})| + 1, 0, i, h)$
  **end**
**end**
**for** $z \in Z$ *in lexicographically ascending order based on the key* $(rank(v), pos, task)$ **do**
  $i \leftarrow z.anchor, h \leftarrow z.haplotype, v \leftarrow z.v, wt \leftarrow weight(M[i])$
  **if** $z.task = 0$ **then**
    $g_1 \leftarrow M[i].x + dist2begin(v,h) + D(v, M[i].v) + M[i].c - 2$
    **if** $h \in haps(M[i].v)$ **then**
      $C(i,h) \leftarrow \max(C_{l2r}(i),\ C(i,h),\ wt + \mathcal{T}_h.\mathsf{RMQ}(0, M[i].c) - g_1)$
      $C''(i) \leftarrow \max(C''(i), C(i,h))$
    **else**
      $C_{l2r}(i) \leftarrow \max(C_{l2r}(i),\ wt + \mathcal{T}_h.\mathsf{RMQ}(0, M[i].c) - g_1 - \gamma)$
  **else if** $z.task = 1$ **then**
    $C(i,h) \leftarrow \max(C''(i) - \gamma,\ C(i,h))$
  **else if** $z.task = 2$ **then**
    $g_2 \leftarrow M[i].y + dist2begin(v,h) + M[i].d$
    $\mathcal{T}_h.\mathsf{update}(M[i].d,\ C(i,h) + g_2)$
**end**

Table S1: We show the true recombinations between haplotypes and the recombinations observed in co-linear chains using recombination penalty $\gamma = 10^4$. We show this data for 45 query sequences that were generated using 0.1% substitution rate. 'S' indicates the start and 'E' indicates the end.

| Query | True haplotype switches | Haplotype switches in chains | F1-score |
|---|---|---|---|
| 1 | >S>HG01123#1>HG02717#1>HG005#1 >HG01358#1>HG01978#2>E | >S>HG02080#2>HG00621#2>HG01978#2>E | 0.200000 |
| 2 | >S>HG002#1>HG01358#2>HG00733#1 >HG01258#1>HG01361#2>E | >S>HG002#2>HG01258#1>HG01361#2>E | 0.400000 |
| 3 | >S>HG00438#1>HG01258#2>E | >S>HG00438#1>HG01258#2>E | 1.000000 |
| 4 | >S>HG02886#1>HG02717#1>HG01928#1 >HG01106#1>HG01071#2>E | >S>HG02886#1>HG002#1>HG01928#2 >HG01106#1>CHM13#0>E | 0.166667 |
| 5 | >S>HG02148#1>HG02717#2>HG01891#2 >HG01123#1>HG01952#1>E | >S>HG02148#1>HG01071#2>HG01891#2 >HG01952#1>E | 0.333333 |
| 6 | >S>HG02622#2>HG01891#1>NA18906#1>E | >S>HG02622#2>HG01891#1>NA18906#1>E | 1.000000 |
| 7 | >S>NA18906#2>HG01358#1>HG02559#2 >HG01109#1>HG005#2>E | >S>HG01358#1>HG02559#2>HG01109#1 >HG005#2>E | 0.727273 |
| 8 | >S>HG01071#2>HG00438#2>HG01891#2>E | >S>HG01071#2>HG00438#2>HG01891#2>E | 1.000000 |
| 9 | >S>HG01175#2>HG01258#1>HG01891#2>E | >S>HG01175#2>HG01258#1>HG01891#2>E | 1.000000 |
| 10 | >S>HG02622#1>HG01952#2>HG01106#1 >HG01358#2>HG00438#1>E | >S>HG02622#1>HG01952#2>HG01106#1 >HG01358#2>HG00438#1>E | 1.000000 |
| 11 | >S>HG00438#1>HG01891#2>E | >S>HG00438#1>HG01891#2>E | 1.000000 |
| 12 | >S>HG02717#2>HG00621#2>HG02723#1>E | >S>HG02717#2>HG00621#2>HG02723#1>E | 1.000000 |
| 13 | >S>HG03098#2>HG02257#2>HG03540#1>E | >S>HG03098#2>HG02257#2>HG03540#1>E | 1.000000 |
| 14 | >S>HG02622#2>HG02559#2>HG00741#1>E | >S>HG02622#2>HG02559#2>HG00741#1>E | 1.000000 |
| 15 | >S>HG01175#2>HG03540#1>E | >S>HG01175#2>HG00735#1>HG03540#1>E | 0.571429 |
| 16 | >S>HG01071#2>HG00673#1>E | >S>HG01071#2>HG00673#1>E | 1.000000 |
| 17 | >S>HG01109#2>HG03516#1>HG02886#1>E | >S>HG01109#2>HG03516#1>HG02886#1>E | 1.000000 |
| 18 | >S>HG03098#2>HG01891#1>E | >S>HG03098#2>HG01891#1>E | 1.000000 |
| 19 | >S>HG01952#1>HG01978#2>HG02886#1 >HG002#1>HG01071#2>E | >S>HG01952#1>HG02886#1>HG005#2 >HG01071#2>E | 0.363636 |
| 20 | >S>HG01952#1>HG002#1>HG002#2 >HG00438#1>HG00733#1>E | >S>HG01952#1>HG002#1>HG002#2 >HG00438#1>CHM13#0>E | 0.666667 |
| 21 | >S>HG02717#2>HG02622#2>HG01952#2>E | >S>HG02717#2>HG02622#2>HG01952#2>E | 1.000000 |
| 22 | >S>HG01978#1>HG03098#2>HG02486#2>E | >S>HG01978#1>HG03098#2>HG02486#2>E | 1.000000 |
| 23 | >S>HG01106#2>HG01258#2>E | >S>HG01106#2>HG01258#2>E | 1.000000 |
| 24 | >S>HG00741#1>HG02886#1>HG01361#1>E | >S>HG00741#1>HG02886#1>HG01361#1>E | 1.000000 |
| 25 | >S>HG02080#1>HG03540#2>HG01106#2 >HG02559#1>HG01358#2>E | >S>HG03540#2>HG01106#2>HG00735#1 >HG005#1>HG01358#2>E | 0.333333 |
| 26 | >S>HG01928#1>HG00735#1>E | >S>HG01928#1>HG00735#1>E | 1.000000 |
| 27 | >S>HG01106#2>HG02717#1>E | >S>HG01106#2>HG00621#2>HG02717#1>E | 0.571429 |
| 28 | >S>HG00735#1>HG01258#1>HG02257#2 >HG00673#1>HG01928#1>E | >S>HG00438#2>HG02257#2>HG00621#2 >CHM13#0>HG01928#1>E | 0.166667 |
| 29 | >S>HG01361#1>HG02723#1>CHM13#0>E | >S>HG01361#1>HG02723#1>CHM13#0>E | 1.000000 |
| 30 | >S>HG02886#2>HG01106#2>HG01978#1 >HG02109#2>HG02148#1>E | >S>HG02886#2>HG01978#1>HG00735#1E >HG02148#1>E | 0.363636 |
| 31 | >S>HG01361#2>HG01109#2>HG01891#1>E | >S>HG01361#2>HG01109#2>HG01891#1>E | 1.000000 |
| 32 | >S>HG02148#1>HG02559#1>E | >S>HG02148#1>HG02559#1>E | 1.000000 |
| 33 | >S>HG00733#1>HG02723#1>HG01175#2 >HG02148#1>HG01123#1>E | >S>HG02109#2>HG02723#1>HG01175#2 >HG02148#1>HG01123#1>E | 0.666667 |
| 34 | >S>HG00673#1>NA18906#1>HG002#1>E | >S>HG00673#1>NA18906#1>HG002#1>E | 1.000000 |
| 35 | >S>HG00621#2>HG005#1>E | >S>HG00621#2>HG005#1>E | 1.000000 |
| 36 | >S>HG00621#2>HG02080#1>HG01358#2>HG03540#1 >HG01361#2>E | >S>HG02109#2>HG01358#2>HG01258#1 >HG01071#2>HG01361#2>E | 0.166667 |
| 37 | >S>HG02257#2>HG005#2>E | >S>HG02257#2>HG005#2>E | 1.000000 |
| 38 | >S>HG03516#1>HG01928#1>HG01952#2 >HG03098#2>HG02559#1>E | >S>HG03098#2>HG01928#1>HG01952#2 >CHM13#0>E | 0.181818 |
| 39 | >S>HG01361#2>HG00438#2>E | >S>HG01361#2>HG00438#2>E | 1.000000 |
| 40 | >S>HG02559#2>HG03098#2>E | >S>HG02559#2>HG03098#2>E | 1.000000 |
| 41 | >S>HG01123#1>HG02257#1>HG01928#2>E | >S>HG01123#1>HG02257#1>HG01928#2>E | 1.000000 |
| 42 | >S>HG01123#2>HG01123#1>E | >S>HG01123#2>HG00735#1>HG01123#1>E | 0.571429 |
| 43 | >S>HG01071#2>HG01928#2>E | >S>HG01071#2>HG01928#2>E | 1.000000 |
| 44 | >S>HG00438#1>HG03540#1>HG02717#2 >HG00735#1>HG02622#2>E | >S>HG02109#2>HG00621#1>HG02622#1 >HG02080#2>HG00735#1>HG02622#2>E | 0.307692 |
| 45 | >S>HG03098#2>HG00621#1>HG01358#1>E | >S>HG03098#2>HG00621#1>HG01358#1>E | 1.000000 |

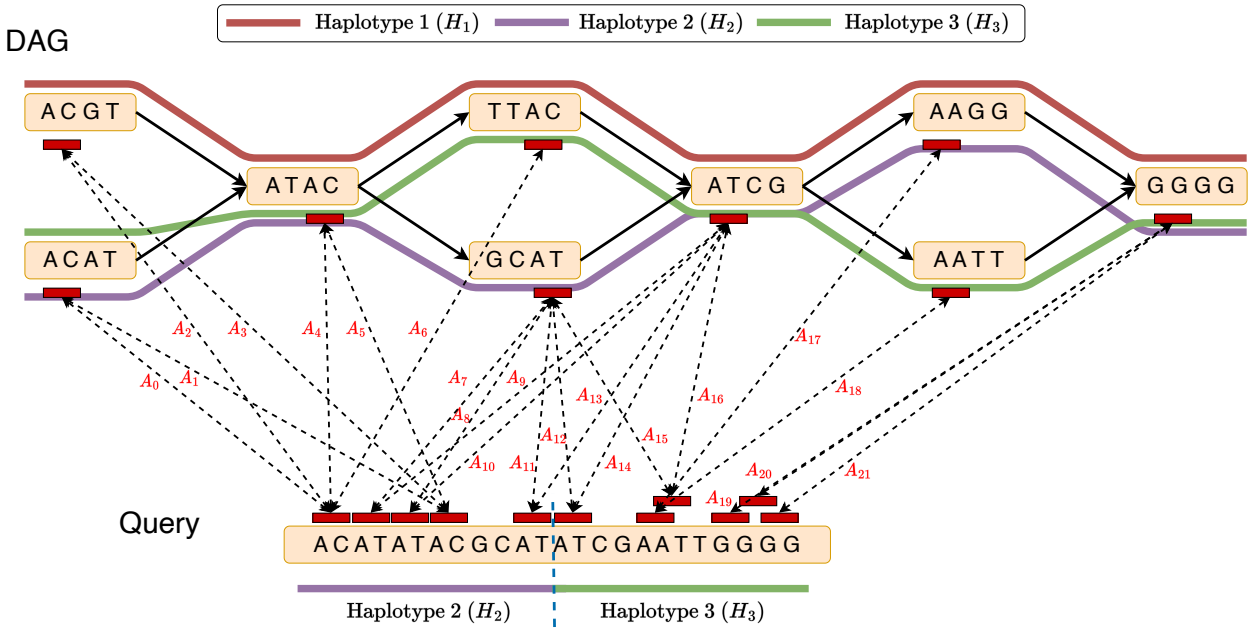
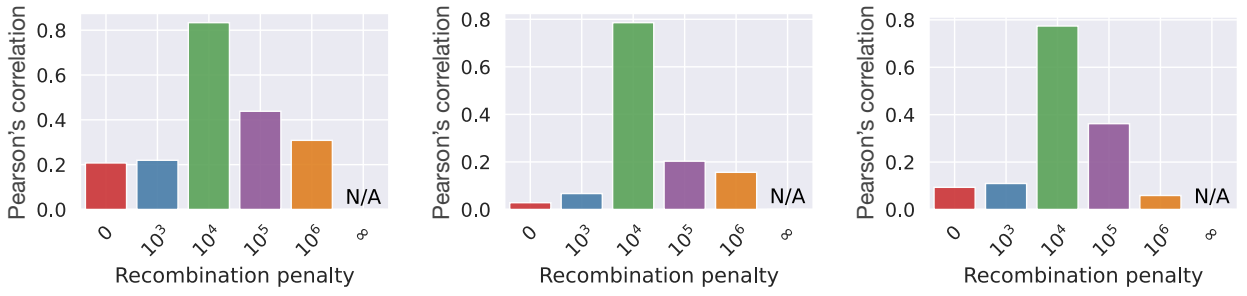
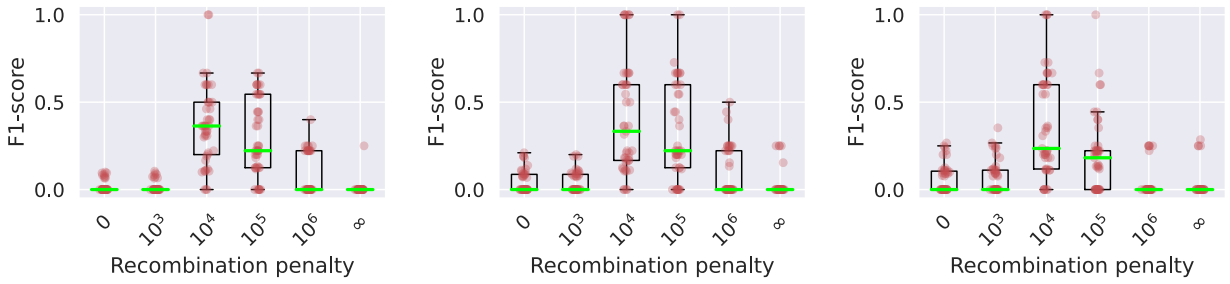



Fig. S1: An example where the query sequence is simulated as a mosaic of two reference haplotypes without substitution error. We use this example to demonstrate that the highest scoring chain produced by the haplotype-agnostic chaining algorithm may be wrong even when there are no substitution errors in the query sequence. The anchors are computed using $(w, k)$ minimizers such that $w = 3, k = 2$. Observe that the true chain includes six anchors $A_0, A_5, A_{11}, A_{14}, A_{18}, A_{19}$ with a switch from haplotype $H_2$ to haplotype $H_3$. Without recombination penalty, the algorithm may output the chain $A_2, A_5, A_{11}, A_{14}, A_{18}, A_{19}$ which involves an incorrect haplotype switch from $H_1$ to $H_2$.



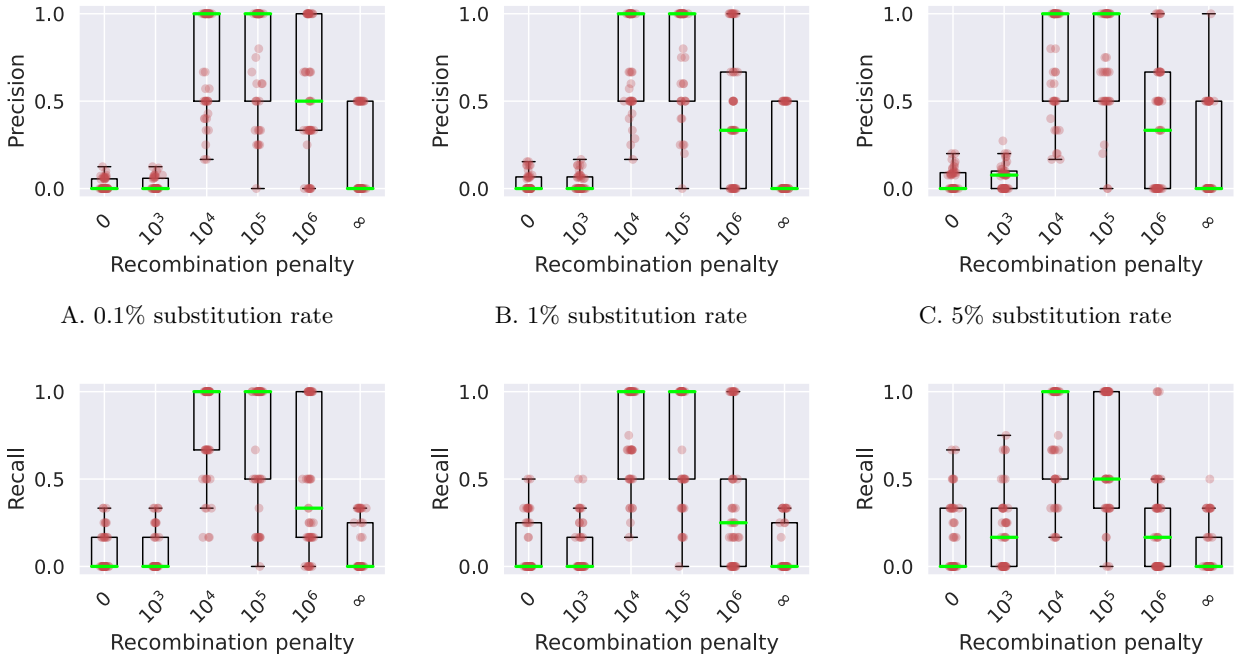

A. 0.1% substitution rate B. 1% substitution rate C. 5% substitution rate

Fig. S2: Pearson's correlation between the number of recombinations in Minichain's output chain and the true count. We evaluated the performance by using different substitution rates and recombination penalties. In this experiment, we simulated query sequences with the number of recombination events ranging from 3 to 9.




A. 0.1% substitution rate B. 1% substitution rate C. 5% substitution rate

Fig. S3: Box plots show the levels of consistency between the haplotype recombination pairs in Minichain's output chain and the ground-truth using three different sets of simulated MHC sequences with substitution rates (a) 0.1%, (b) 1%, and (c) 5%. We tested using different recombination penalties. Each red dot in the plots corresponds to a query sequence. The median values are highlighted in green. In this experiment, we simulated query sequences with the number of recombination events ranging from 3 to 9.




A. 0.1% substitution rate B. 1% substitution rate C. 5% substitution rate

D. 0.1% substitution rate E. 1% substitution rate F. 5% substitution rate

Fig. S4: Box plots show the precision and recall between the haplotype recombination pairs in Minichain's output chain and the ground-truth using three different sets of simulated MHC sequences with substitution rates (a), (d) 0.1%, (b), (e) 1%, and (c), (f) 5%. We tested using different recombination penalties. Each red dot in the plots corresponds to a query sequence. The median values are highlighted in green.