Supplementary Materials: Simultaneous structural variation discovery among multiple paired-end sequenced genomes

Fereydoun Hormozdiari [1,†], Iman Hajirasouliha [1,†], Andrew McPherson [1], Evan E. Eichler [2], S. Cenk Sahinalp[1]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada

[2] Department of Genome Sciences, University of Washington, and Howard Hughes Medical Institute, Seattle, WA, USA

[†] Joint First Authors.

Supplemental materials

Proof of Theorem 1

Red-Black-Assignment-F2 can be approximated within a factor of $1 + \frac{\omega_{\max}}{\omega_{\min}}$.

Proof. We remind the reader that the instance of Red-Black-Assignment-F2 problem is denoted by H and a maximal matching in H is denoted by M. $R = \{r_1, r_2, \cdots, r_p\}$ is the set of red edges and $B = \{b_1, b_2, \cdots, b_q\}$ is the set of black edges in M.

Our algorithm first probes all the edges in R (the set of red edges in the maximal matching) and assigns them to one of their vertices. Each red edge $r_i \in R$ is from one of the following categories:

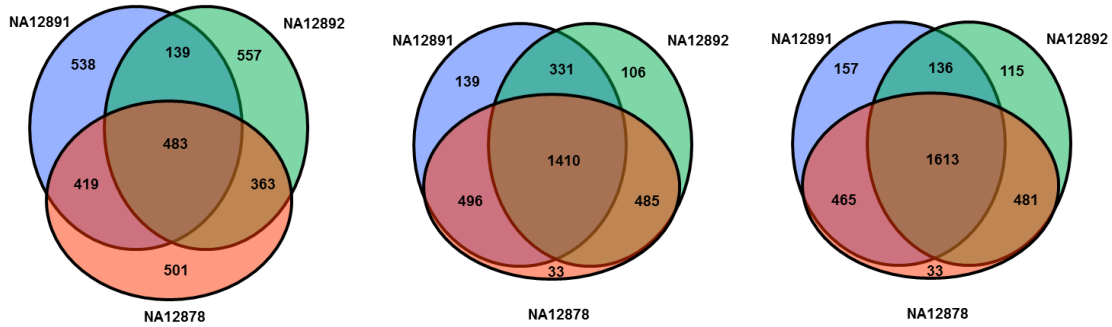
- *There exists a black edge specific to r_i in H:* in other words, this black edge shares a vertex with r_i but does not share a vertex with any other red edge in R. In this case, the algorithm simply orients both r_i and the above-mentioned black edge to this shared vertex.
- *r_i does not share a vertex with a black edge in H:* In this case the algorithm orients r_i arbitrarily.
- *Each black edge sharing a vertex with r_i has its other vertex shared by another red edge $r_j \in R$:* Let $R' \subseteq R$ be the set of red edges which share a vertex with a black edge - not specific to any red edge. We construct a *new* graph $H^{R'}$ as follows: corresponding to each edge $r'_j = (x'_j, y'_j)$ in R' set up a *vertex* ρ'_j in $H^{R'}$. For each pair of vertices ρ'_k and ρ'_ℓ in $H^{R'}$ and for each black edge in H which share vertices with both r'_k and r'_ℓ, set up an edge $e'_{k,\ell}$ connecting ρ'_k and ρ'_ℓ. Note that $H^{R'}$ is not necessarily a simple graph. Suppose $H^{R'}$ has t connected components denoted by $C_1, \cdots, C_t$. For each C_i, we first orient its edges such that each vertex has an *indegree* at least 1. Note that such an orientation can always be discovered via a Depth-first search (DFS) algorithm, unless C_i is a (simple) tree in which exactly one vertex (the root of the DFS) would have indegree equal to zero (i.e. no edges terminating at it). WLOG, let the direction of the edge $e'_{k,\ell}$ be from ρ'_k to ρ'_ℓ. We orient the black edge $e'_{k,\ell}$ towards its vertex (say x_ℓ), which is shared by r'_ℓ. The edge r'_ℓ will also be oriented to x_ℓ and thus x_ℓ will be multicolor. This guarantees that all but one of the red edges in R' will be oriented towards a vertex, also oriented by a black edge.

We will use a similar strategy for the set of black edges in the matching and finally orient all the remaining edges in H arbitrarily. This strategy will guarantee that even if the optimal solution covers an edge $e_M \in M$ with a multicolor vertex and does not pick the other vertex of e_M (i.e. incurring a cost of only $\omega_{\min}$), e_M can be covered with a cost of at most $\omega_{\max} + \omega_{\min}$ by selecting both of its vertices - which will ensure at least one of its vertices will be multicolor.

If the optimal solution covers e_M with a single colored vertex, our strategy will cover it with a cost of at most $2 \cdot \omega_{\max}$, providing us a $1 + \omega_{\max}/\omega_{\min}$ approximation factor.

□

CEU Trio *Alu* insertion results



(a) ISV&M (b) SSC (c) SSC-W

Figure 1: Figures (a), (b) and (c) detail the number of common and de novo events in each genome for the ISV&M, SSC and SSC-W respectively for the CEU trio.