RIFRAF: a frame-resolving consensus algorithm
Supporting information

Kemal Eren, Ben Murrell

November 30, 2017

1 NL4-3 and references

The distribution of estimated error rates of the NL4-3 reads appears in Figure S1.

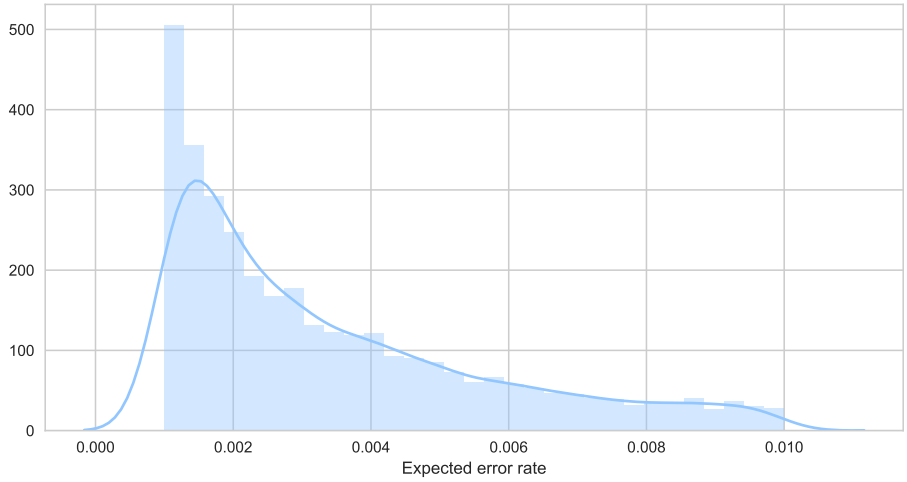


Figure S1: QV-derived error rates of the NL4-3 sequences used.

The phylogenetic tree of the reference database appears in Figure S2.

2 Length of template experiment

As a proxy for varying amplicon lengths, we sampled prefixes of varying lengths from our NL4-3 sequence dataset. This protocol is valid because SMRT sequencing does not have a positional bias (see Figure S8D in [3]), so the distribution of errors in the prefix should match the overall errors. However, these short reads

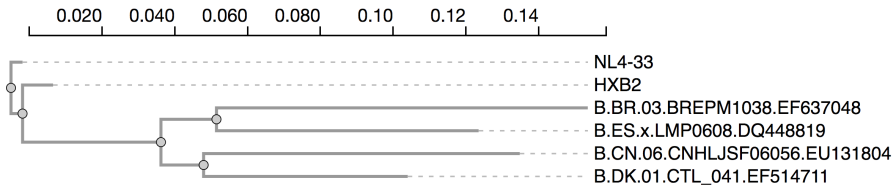
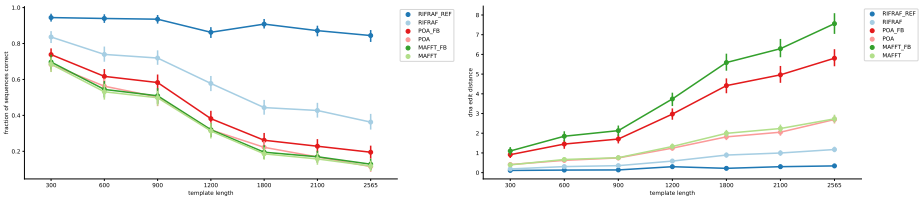


Figure S2: Phylogeny of references used by $\texttt{RIFRAF}_{\texttt{ref}}$. The references were aligned with `MAFFT` [1, 2], the phylogeny inferred by FastTree [4, 5], and visualized with PhyloTree.js (`https://github.com/veg/phylotree.js`).

are of lower overall quality than true short CCS reads, which have the benefit of more subreads in the ZMW read.

The experiment was run for 1,000 trials of clusters of size $N = 3$. The fraction of correct consensus sequences appears in Figure S3a. Although all methods degrade in accuracy for longer sequences, $\texttt{RIFRAF}_{\texttt{ref}}$ degrades much more slowly, getting at least 80% perfectly correct even for the full-length amplicon. For this number of sequences, $\texttt{RIFRAF}_{\texttt{ref}}$ beats even the other methods with frame correction. As expected, frame correction keeps all three methods in-frame (Figure S4a). The amino acid edit distance for frame correction methods alone appears in Figure S4b for clarity, showing that $\texttt{RIFRAF}_{\texttt{ref}}$'s frame-corrected amino acid sequences contain an average of less than one error even for the full-length amplicon, whereas the accuracy of other methods degrades from one error all the way to three errors for `POA_FB` and four errors for `MAFFT_FB`.



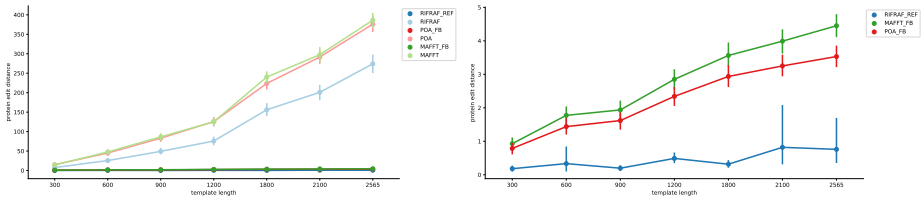
(a) Fraction of correct sequences versus template length.

(b) DNA edit distance versus template length.

Figure S3: Results on DNA sequences for varying sequence length. The left figure shows the fraction of correct sequences; the right figure shows mean edit distance between the consensus and the template.

3 True indel experiments

Single indels were simulated in both homopolymer (defined as four or more identical bases in a row) and non-homopolymer regions of NL4-3 in the follow-



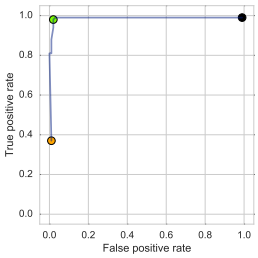
(a) Protein edit distance versus template length; all methods.

(b) Protein edit distance versus template length; frame correction only

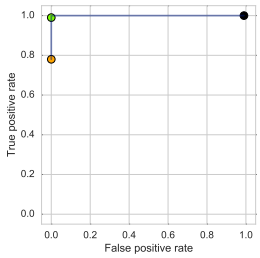
Figure S4: Same results as Fig. S3, but for the translated protein sequences. The fraction of correct sequences is not reproduced, since those figures are identical. The left figure show results for all methods. The right figure shows the same data, zoomed in on the details of the frame-corrected results.

ing manner. First, a region was sampled uniformly from all matching regions; i.e. a homopolymer region is chosen at random from all homopolymer regions. Insertions or deletions in homopolymer regions were simulated by inserting or deleting a base. Deletions to non-homopolymer regions were simulated by removing a single random base from that region. Insertions in non-homopolymer regions were simulated by choosing a non-homopolymer position and inserting a random base either before or after it.

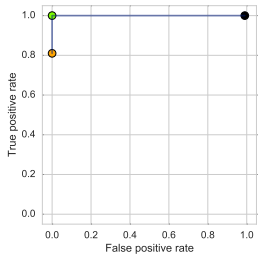
Each read has a small probability p of not being modified to match the new template, where p is proportional to the read's estimated error rate and is calculated as the mean of the Phred scores after converting them to error probabilities. Those reads that were modified were pairwise aligned to the template and the matching base was inserted or deleted from the correct position. Quality scores for insertions were drawn from a uniform distribution between the two Phred scores on either side of the insertion.



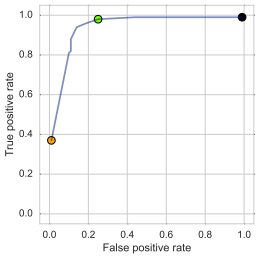
(a) $N = 3$, non-HP insertion



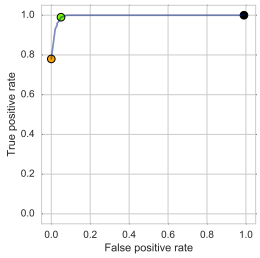
(b) $N = 5$, non-HP insertion



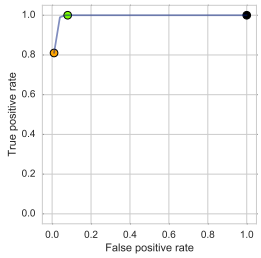
(c) $N = 10$, non-HP insertion



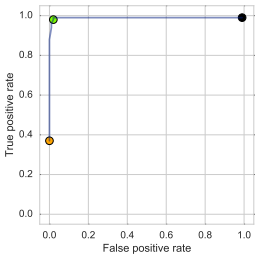
(d) $N = 3$, HP insertion



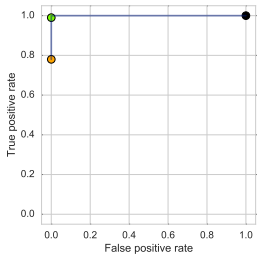
(e) $N = 5$, HP insertion



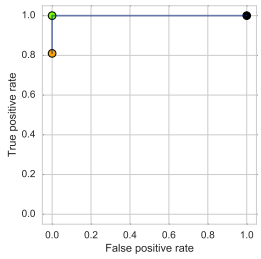
(f) $N = 10$, HP insertion



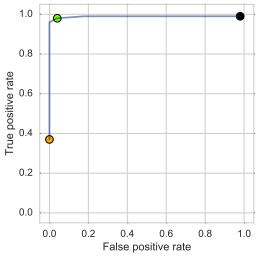
(g) $N = 3$, non-HP deletion



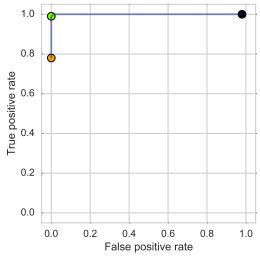
(h) $N = 5$, non-HP deletion



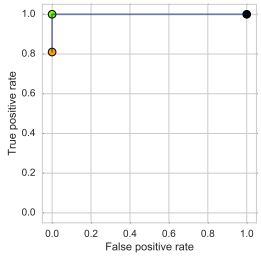
(i) $N = 10$, non-HP deletion



(j) $N = 3$, HP deletion

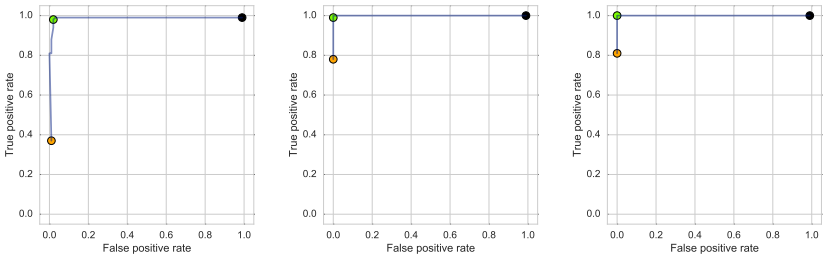


(k) $N = 5$, HP deletion

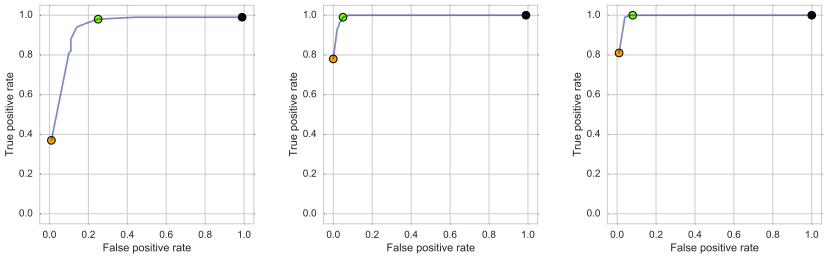


(l) $N = 10$, HP deletion

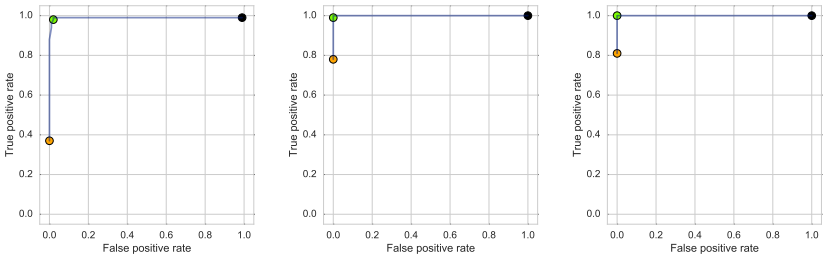
Figure S5: ROC curves for true indel experiments, with max indel penalty multiplier m from 0 to 12. Orange points denote runs without a reference; the rest use HXB2 as a reference. The green point corresponds to a max indel penalty multiplier $m = 6$.



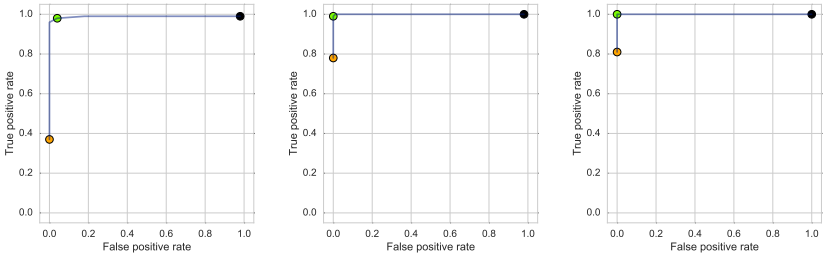
(a) $N = 3$, non-HP insertion (b) $N = 5$, non-HP insertion (c) $N = 10$, non-HP insertion



(d) $N = 3$, HP insertion (e) $N = 5$, HP insertion (f) $N = 10$, HP insertion



(g) $N = 3$, non-HP deletion (h) $N = 5$, non-HP deletion (i) $N = 10$, non-HP deletion



(j) $N = 3$, HP deletion (k) $N = 5$, HP deletion (l) $N = 10$, HP deletion

Figure S6: The same results as in Figure S5, except using the more distant sequence B.BR as a reference.

References

[1] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

[2] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.

[3] Melissa Laird Smith, Ben Murrell, Kemal Eren, Caroline Ignacio, Elise Landais, Steven Weaver, Pham Phung, Colleen Ludka, Lance Hepler, Gemma Caballero, Tristan Pollner, Yan Guo, Douglas Richman, Pascal Poignard, Ellen E. Paxinos, Sergei L. Kosakovsky Pond, and Davey M. Smith. Rapid Sequencing of Complete env Genes from Primary HIV-1 Samples. *Virus Evolution*, 2(2), 2016.

[4] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[5] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.