

Introduction

FILLME

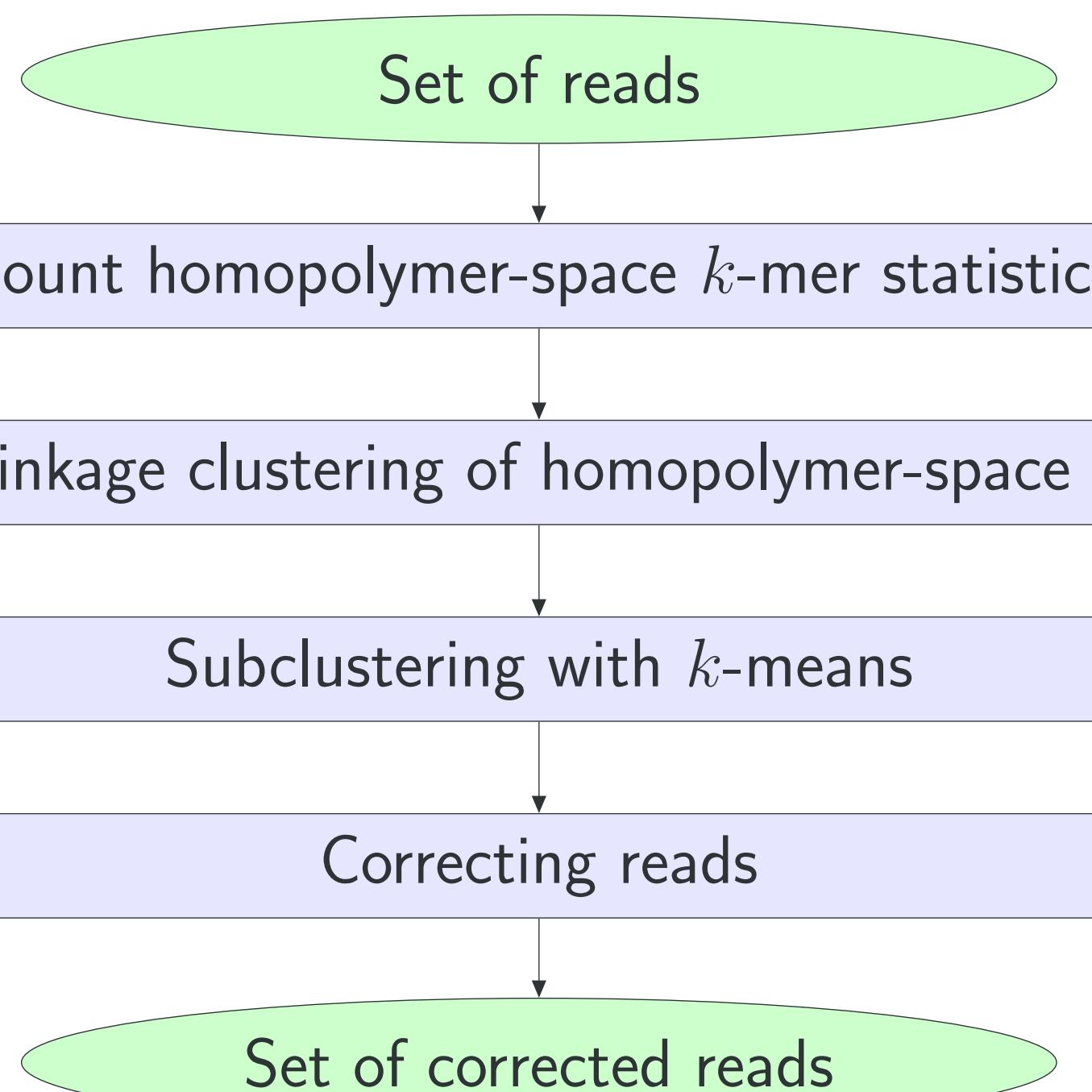
Notation

Let \mathcal{N} denote the nucleotide alphabet $\{A, C, G, T\}$. By definition, put $\mathcal{H} = \mathcal{N} \times \mathbb{N}^+$.

We call an element of the alphabet \mathcal{H} a *homopolymer run*, and an element of \mathcal{H}^k a *homopolymer-space k-mer*.

We use $x[k]$ to denote k -th element of a sequence, using zero-based indexing; $x[k \dots l]$ to denote a subsequence $x[k]x[k+1] \dots x[l]$; finally, length of x is denoted by $|x|$. Distance between $i, j \in \mathcal{H}^k$ is defined as the minimum number of 1-base insertions/deletions/mismatches needed to align common part of i and j . It is denoted by $\text{dist}(i, j)$.

HammerIT workflow



Pairwise distance calculation

We use 5-base lookahead to compute distance between k -mers in homopolymer-space. Helper table stores precomputed values of

$$H_k : \mathcal{N}^k \times \mathcal{N}^k \rightarrow \{\text{Insertion, Deletion, Mismatch}\}, \quad k = 1, 2, 3, 4, 5.$$

The chosen value of 5 is a trade-off between accuracy and speed.

(Mapped reads from B7-295.bam dataset were used for training.)

$H_5(\text{AGTAC}, \text{TGTAC}) = \text{Mismatch}$

$H_5(\text{CTGTC}, \text{TGTCA}) = \text{Deletion}$

$H_3(\text{TCG}, \text{ATG}) = \text{Mismatch}$

$H_2(\text{CG}, \text{TG}) = \text{Mismatch}$

Algorithm

Input: $x, y \in \mathcal{N}^+$ — homopolymer-space k -mers in nucleotide alphabet

Output: dist — approximate distance between x and y

$\text{pos}.x \leftarrow 0; \text{pos}.y \leftarrow 0; \text{dist} \leftarrow 0;$

while $\text{pos}.x < |x|$ and $\text{pos}.y < |y|$ **do**

if $x[\text{pos}.x] = y[\text{pos}.y]$ **then**

$\text{pos}.x \leftarrow \text{pos}.x + 1; \text{pos}.y \leftarrow \text{pos}.y + 1;$

else

$k \leftarrow \min(5, |x| - \text{pos}.x, |y| - \text{pos}.y);$

 adjust $\text{pos}.x$ and $\text{pos}.y$ according to

$$H_k[x[\text{pos}.x \dots \text{pos}.x + k - 1], y[\text{pos}.y \dots \text{pos}.y + k - 1]];$$

$\text{dist} \leftarrow \text{dist} + 1;$

IonTorrent error profile

IonTorrent error profile

Corrected flow signal intensities are available in BAM files produced by versions of Ion Torrent Suite prior to 3.4. Called homopolymer length is obtained as corrected flow signal intensity rounded to the nearest integer.

We have studied flow signal intensity distributions around insertion/deletion sites. File B7-295.bam, downloaded from Ion Community website, contained 4.6M insertions, 5.0M deletions, and 1.5M mismatches.

Overwhelming majority of errors turned out to be insertions/deletions of length 1, occurring when flow signal intensity is approximately halfway between two adjacent integers.

Figure 1: Flow signal intensities at insertion sites

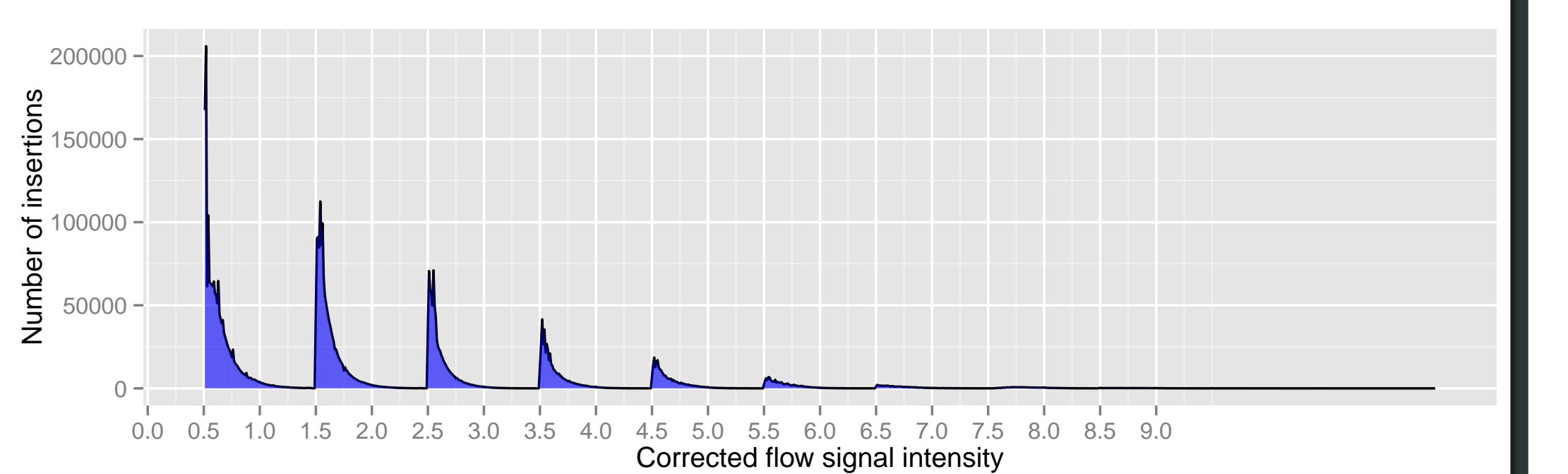
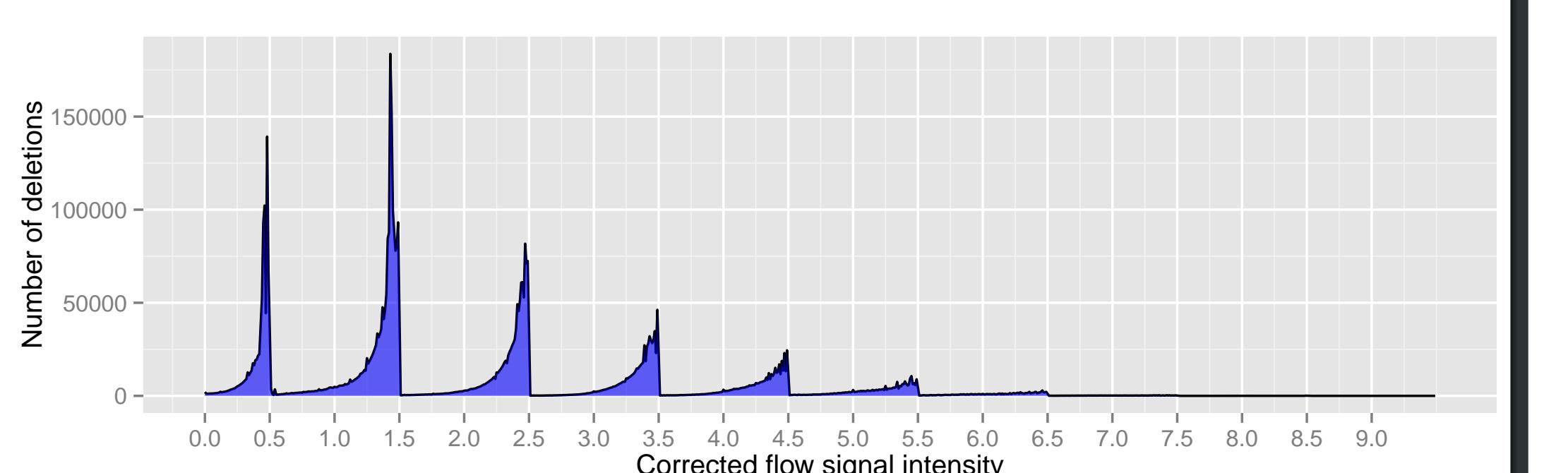


Figure 2: Flow signal intensities at deletion sites



More detailed analysis of errors in IonTorrent data can be found in the article "Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data" (PLoS Comput Biol 9(4))

Error reduction results

We evaluated HammerIT on 6 publicly available datasets, using the same pipeline as the authors of the recently published article "Updating benchtop sequencing performance comparison" (Nature Biotechnology, v. 31, no. 4). In that article, error rate in four Ion Torrent datasets has been assessed. We used the same data plus two extra datasets from 314v2 chip, which recently became available on Ion Community Portal.

Indel/mismatch error rates were calculated for uniquely mapped reads before and after correction. For each dataset, correction was done in two ways. In the first setup, trimming was done for read ends that couldn't be corrected due to lack of good k -mers, while in the second one such read ends were preserved in the output. Relative change in read coverage after correction stayed within 0.4% in all cases.

Figure 3: Error rates before and after correction

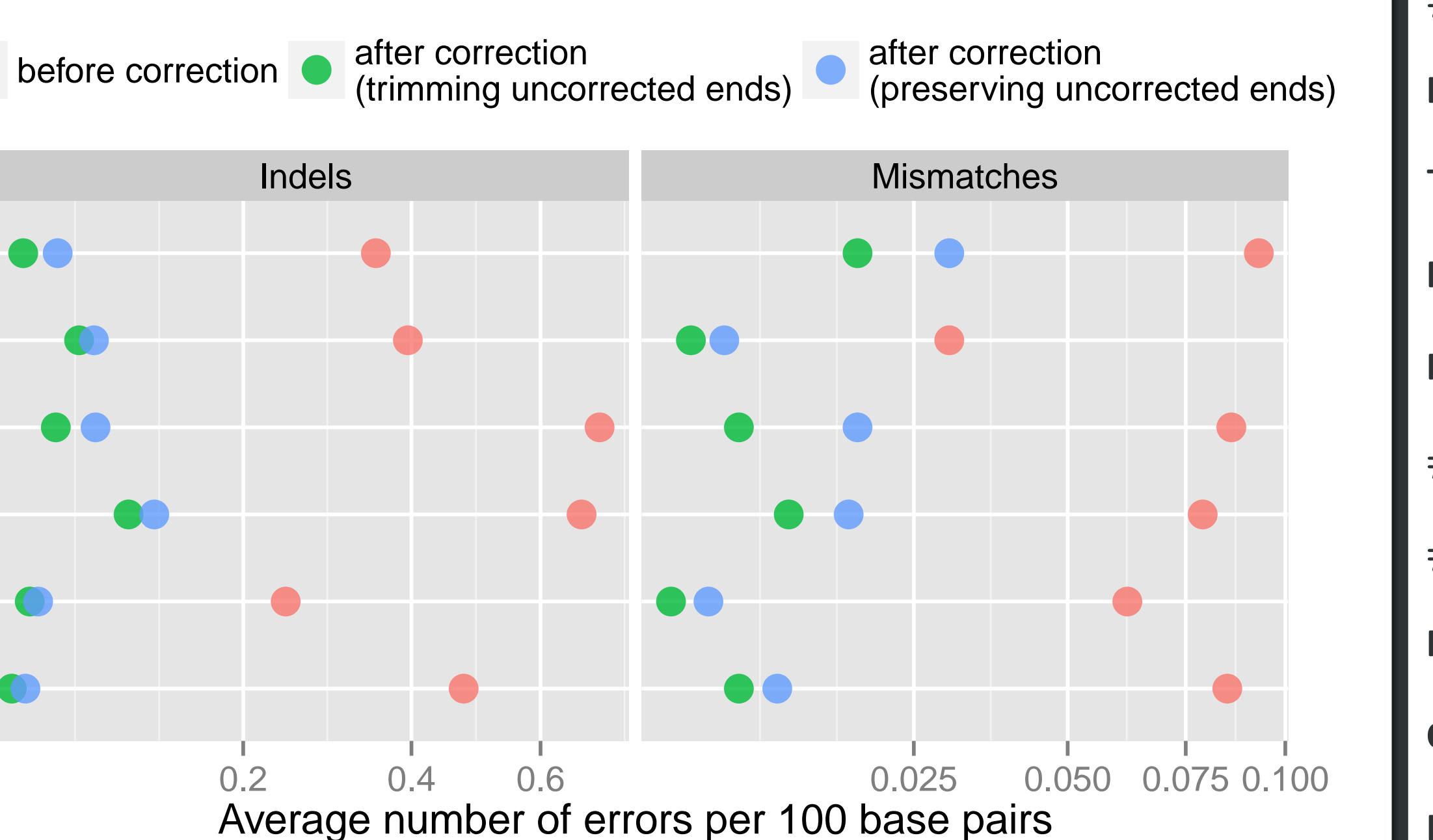


Figure 4: Error reduction by read position for Sakai 400bp reads



Assembly results

We have assembled *E.coli O157 H7 Sakai* 400bp reads before and after correction with MIRA 3.4.1.1 and SPAdes 2.4.0. For the MIRA assembler, the reads were subsampled to get ~40x coverage as recommended by the authors of the original benchmark, whereas SPAdes assembler operated on the full dataset (~200x).

Length of the reference genome is 5498450 bp.

Table 1: Assembly results, *E.coli* Sakai dataset (400bp). Contigs of length ≥ 500 are used.

	MIRA 3.4.1.1	SPAdes 2.4.0		
	uncorrected	corrected	uncorrected	corrected
# contigs	178	197	266	242
Largest contig	378520	378249	316904	374932
Total length	5510229	5517674	5322223	5320566
N50	147710	124917	111277	146551
N75	76533	70728	44319	65938
# misassemblies	23	19	0	2
# local misassemblies	17	20	8	11
Misassembled contigs length	842594	567793	0	32563
Genome fraction (%)	97.241	97.441	93.880	93.966
Duplication ratio	1.011	1.015	1.001	1.001
# mismatches per 100 kbp	12.68	22.25	3.58	4.70
# indels per 100 kbp	9.85	6.64	6.21	5.88

Command-line parameters used for assembly

MIRA 3.4.1.1 --job=denovo,genome,accurate,iontor --notraceinfo IONTOR_SETTINGS

--ASSEMBLY:mprc=100

SPAdes 2.4.0 --only-assembler -k 21,33,55,77,99

Acknowledgements

FILLME

Error correction algorithm

```

for each read from the dataset do
  (Producing contiguous corrected parts of read)
  for each homopolymer-space kmer from the read do
    center ← center of the cluster to which kmer belongs;
    if center quality is more than user-specified threshold and
      center bases agree with the previous "good" center then
        include center into consensus score calculation;
    else
      yield new corrected part from current consensus;
      trim homopolymer runs with low consensus score from ends;
      reset consensus table and start a new part;
  (Combining corrected parts)
  while there are two or more parts do
    curr ← first part; next ← second part;
    align last 8 homopolymer runs of curr against the read;
    align first 8 homopolymer runs of next against the read;
    if there is a gap on the read between the two parts then
      copy read homopolymer runs as is;
    else
      select homopolymer runs with higher consensus score
        from the intersection of the two parts;
      replace curr and next with the combined part;
  (Optionally, attaching uncorrected end)
  align last 8 runs of the last chunk against the read sequence;
  append read homopolymer runs after the last aligned run.

```