Read mapping tool for AB SOLiD data

{Marta.Girdea, Laurent.Noe, Gregory.Kucherov}@lifl.fr

lifL Université Lille1 Sciences et Technologies INRIA

Contribution

We developed a sensitive tool for mapping AB SOLiD reads, that makes use of the **AB SOLiD color code properties** and **read qualities** to ensure a **fast**, **"base intelligent"**, **indel capable** identification, alignment and mapping.

AB SOLiD color space

AB SOLiD sequencers [1] produce reads encoded in a 4 color space, whose error-correcting properties help to distinguish SNPs from reading errors.

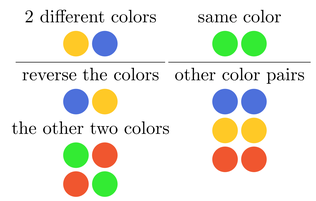
Di-nucleotide color table

1st base \ 2nd base	A	C	G	T
A	AA	AC	AG	AT
C	CC	CA	CT	CG
G	GG	GT	GA	GC
T	TT	TG	TC	TA

Color sequence example

	A	C	G	A	G	G	T
A	C	G	A	G	G	T	
C	A	T	C	T	T	G	
G	T	A	G	A	A	C	
T	G	C	T	C	C	A	

SNP detection



SNP vs. reading error

Approach

All the data processing is performed in the **color space**. The reference genome is translated into colors and indexed accordingly.

Step 1: Filtration (a) For each read, candidate mapping positions are identified using specially designed **seeds**.

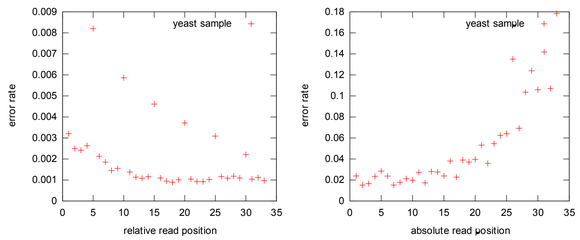
Step 2: Filtration (b) A **fast SIMD bandwidth alignment algorithm** discards candidate mapping positions where the corresponding reference fragment does not show sufficient similarity with the read.

Step 3: Alignment A base-intelligent, gapped alignment algorithm matches the read to the corresponding alignment fragment. **The best N candidates are stored for each read**.

Step 4: Mapping Best scoring reads are mapped first, and used to decide the next mappings.

Reading errors distribution

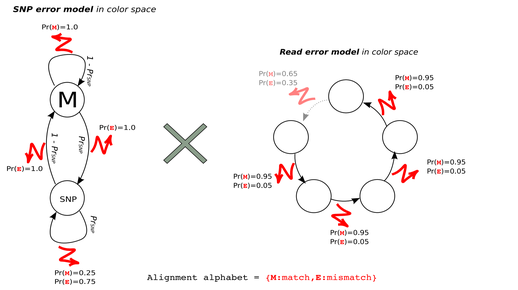
In the SOLiD technology, di-nucleotides are read in cycles of 5 [1]. Data shows that any bias, such as the chance of reading error is propagated in periods of 5.



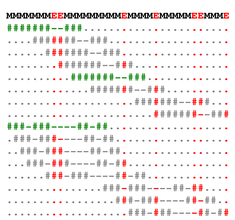
Additionally, the second graph shows the reading error frequency on each position of the read: errors are more likely to occur at the end of the read, especially on the last 10 colors.

Filtration (a): Seeds

For finding candidate mapping positions, we use **Iedera** [2] to design efficient **seeds**, based on a **model that reflects the reading error distribution** observed on SOLiD reads.



Iedera can associate to each seed the list of relevant positions on the read to which the search should be restricted, optimized according to the seed patterns.



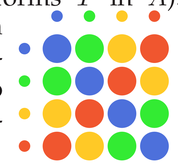
Filtration (b): SIMD filter

Most false positive hits are detected and eliminated by a fast SIMD bandwidth alignment of the read and the reference that can process several hits in a single run.

Alignment: preamble

To obtain alignments of color sequences that are meaningful in the nucleotide space, color pair alignments must be implicitly interpreted as nucleotide alignments.

Colors can be seen as *transformations* of bases [1] (for example, ● transforms T in A). A color sequence can be seen as a series of successive base transformations. They can be *composed*, to obtain the color with the same effect as the whole sequence.



Properties of color composition [1]: *commutativity*, *associativity*, ● is the *neutral element*, each color is its own *inverse*.

Detecting valid DNA modifications:

Consecutive mutations

$$\begin{array}{ccccccccccc} B_1 & & B_2 & & B_3 & & B_4 & & B_5 \\ & c_1 & & c_2 & & c_3 & & c_4 & \\ B_1 & & B'_2 & & B'_3 & & B'_4 & & B_5 \end{array}$$

$$c_1 \neq c'_1$$
$$c_1 \cdot c_2 \neq c'_1 \cdot c'_2$$
$$c_1 \cdot c_2 \cdot c_3 \neq c'_1 \cdot c'_2 \cdot c'_3$$
$$c_1 \cdot c_2 \cdot c_3 \cdot c_4 = c'_1 \cdot c'_2 \cdot c'_3 \cdot c'_4$$

Consecutive indels

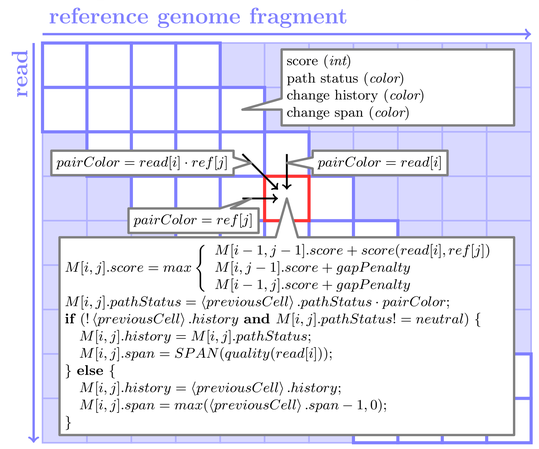
$$\begin{array}{ccccccccccc} B_1 & & B_2 & & B_3 & & B_4 & & B_5 \\ & c_1 & & c_2 & & c_3 & & c_4 & \\ B_1 & & - & & - & & - & & B_5 \end{array}$$

$$c'_1 = c_1 \cdot c_2 \cdot c_3 \cdot c_4$$

Alignment algorithm

Bandwidth alignment, with a configurable number of indels allowed.

The algorithm is based on the classic semi-global sequence alignment approach, enriched with a **limited memory of the previous color mismatches** on each path of the alignment matrix. Unless a color mismatch is "corrected" (followed by other mismatches that will eventually lead to the same base in both nucleotide sequences) within a number of steps, it is considered a reading error.

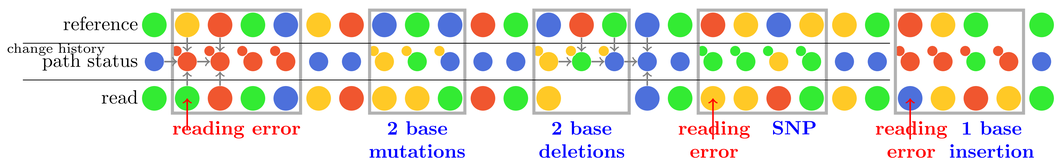


SNP, reading error and indel detection on an alignment path

```
if (M[i, j].span == 0 and M[i, j].history) {
  // Color mismatch not corrected within
  // span: Reading error detected
  // Correction of pathStatus and history
  M[i,j].pathStatus ·= M[i, j].history;
  erase M[i, j].history;
}
```

```
else if (M[i, j].pathStatus == neutral
and M[i, j].history) {
  // SNPs or base indels detected
  // Reset the history
  erase M[i, j].history;
  // Correct score of valid base alignment
  M[i, j].score += reward;
}
```

```
else if (after some color mismatches, M[i, j].pathStatus ==
M[i, j].history) {
  // Reading error + SNPs/indels
  // Correction color sequence, reset history and status
  M[i,j].pathStatus ·= M[i, j].history; erase M[i, j].history;
  // Correct score of valid base alignment
  M[i, j].score += reward;
}
```



Note: span *(ranging between [1..4])*, match scores *([0..3]) and* mismatch penalties *([-3..0]) are stronger for high quality read colors.*

Mapping

The reads are mapped in decreasing order of their score (most "trusted" reads first). When mapping a new read, its score and traceback are adjusted according to a score-weighted multiple alignment of the already mapped reads it overlaps, improving the choice between candidate mapping positions.

References

[1] Applied Biosystems. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction. Methods for Annotating 2 Base Color Encoded Reads in the SOLiD™System 2008

[2] Kucherov, G. and Noé, L. and Roytberg, M. A unifying framework for seed sensitivity and its application to subset seeds *Journal of Bioinformatics and Computational Biology, Springer, 2006*