**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

**APPROACH**

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

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

2. **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.

3. **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.

4. **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.

**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, \( \Delta \) 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, \( \Delta \) is the neutral element, each color is its own inverse.

Detecting valid DNA modifications:

**FILTRATION (A): SEEDS**

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

Idera 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.

**ALGORITHM**

**Bandwidth alignment**, with a configurable number of indels

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.

**REFERENCES**
