Mapping reads from AB Solid sequencers

May 5th 2009
1. Overview of the AB Solid Color Space
   - The color code
   - Properties
   - SNPs vs. reading errors
   - Valid base indels

2. Our read mapping tool
   - Requirements
   - Our approach
Mapping reads from AB Solid sequencers

Overview of the AB Solid Color Space
The AB Solid Color Space: an example

C A G C A T C T T T G T C A T
The AB Solid Color Space: an example

Mapping reads from AB Solid sequencers
The AB Solid Color Space: an example
The AB Solid Color Space: an example

C A G C A T C T T G T C A T
The AB Solid Color Space: an example

C A G C A T C T T G T C A T
The AB Solid Color Space: an example

Mapping reads from AB Solid sequencers
The AB Solid Color Space

The diagram illustrates the mapping of nucleotide pairs in the AB Solid color space. The first base (A, C, G, T) is paired with the second base (A, C, G, T), resulting in a matrix of possible combinations. The colors represent different nucleotide pairs:

- Red: AT
- Green: AC
- Blue: AA
- Orange: GA
- Yellow: CG
- Purple: CA
- Pink: CC
- Brown: TC
- Light Blue: GC
- Light Purple: GT
- Light Green: GG
- Dark Green: AG
- Light Orange: TA
- Orange: TG
- Pink: TT
- Green: CT

This mapping is used to read sequence data from AB Solid sequencers.
Properties

- A pair of bases has the same color regardless of the order:
  \[ \text{color(AG)} = \text{color(GA)} \]

- Given a base, it always results in a different color when paired with a different base:
  \[ \text{color(AG)} \neq \text{color(AC)} \neq \text{color(AT)} \neq \text{color(AA)} \]

- A pair of bases has the same color with its complementary:
  \[ \text{color(AG)} = \text{color(TC)} \]
The AB Solid Color Space: an example (2)

Mapping reads from AB Solid sequencers
SNP detection

- in a color sequence, a base mutation is marked by two adjacent color changes;
- only 3 of the 15 possible pairs correspond to a valid base mutation: the inverse, and both pairs involving the other two colors.
Valid SNPs – example

Mapping reads from AB Solid sequencers
SNP detection – rules

If 2 colors present: e.g.

- Reverse the colors

- Use other 2 colors, both combinations

If 1 color present: e.g.

- Can be the other 3 color pairs:

Mapping reads from AB Solid sequencers
SNP vs reading error

SNP site indicated by 2 adjacent color changes

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>G</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td>C</td>
<td></td>
<td></td>
<td>T</td>
<td></td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Read in color space
Read in base space

Single color change is typically a measurement error

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>C</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>G</th>
<th>G</th>
<th>T</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>G</td>
</tr>
<tr>
<td>Color</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reference in base space
Reference in color space
Read in color space
Read in base space
SNP vs reading error

SNP site indicated by 2 adjacent color changes

Single color change is typically a measurement error

Reference in base space
Reference in color space
Read in color space
Read in base space
**Color “composition” table (1)**

<table>
<thead>
<tr>
<th>1st color</th>
<th>2nd color</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mapping reads from AB Solid sequencers
Color “composition” properties

- commutative,
- associative,
- the color BLUE is the neutral element,
- each color is its own inverse;
- if we associate to each color a 2 bit encoding as follows:

\[
    B = 00, \quad G = 01, \quad Y = 10, \quad R = 11
\]

the color composition operation can be implemented as **XOR**.
Color “composition” table (2): XOR

```
<table>
<thead>
<tr>
<th>1st color</th>
<th>00</th>
<th>01</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00</td>
<td>01</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>01</td>
<td>01</td>
<td>00</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>11</td>
<td>00</td>
<td>01</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>10</td>
<td>01</td>
<td>00</td>
</tr>
</tbody>
</table>
```
Mapping reads from AB Solid sequencers

Our read mapping tool
Related work

Existing software for read mapping

- ZOOM
- ShRiMP
- PASS
- MAQ
- Bowtie
Requirements

- Fast
- Use read qualities
- Allow indels
- Base intelligent substitutions and indels (see next slides for examples)
Undrestanding read qualities

\[ Q = -10 \cdot \log_{10} (P_e) \]  

(0)

where \( P_e \) = error probability.

- \( Q = 40 \Rightarrow P_e = 0.0001 \)
- \( Q = 30 \Rightarrow P_e = 0.001 \)
- \( Q = 20 \Rightarrow P_e = 0.01 \)
- \( Q = 10 \Rightarrow P_e = 0.1 \)
- \( Q = 2 \Rightarrow P_e = 0.63 \)
Example 1: color changes

Not every isolated 1-color change is a reading error

Reference: A C G T A C G T A C G T A C
Read: A C G T A C C A A C G T A C

Reference: A C G T A C G T A C
Read: A C G T A G C A A C G T A C
Example 2: Color Changes

Not every impossible 2-color change is a reading error.

Reference: $\text{ACGTACGTAC}$

Read: $\text{ACGTACACAAACGTAC}$
Example 3: Color Changes

Not every impossible isolated 2-color change is a reading error

Reference: A A A

Read: A A A

Mapping reads from AB Solid sequencers
Base mutation detection: rules

Reference: $B_1 \ B_2 \ B_3 \ B_4 \ B_5$

Read: $B_1 \ B_6 \ B_7 \ B_8 \ B_5$

$c_1 \ c_2 \ c_3 \ c_4$

Properties:
1. $c_1 \neq c_5$
2. $c_1c_2 \neq c_5c_6$
3. $c_1c_2c_3 \neq c_5c_6c_7$
4. $c_1c_2c_3c_4 = c_5c_6c_7c_8$
**Example 4: indels**

Reference: \[ \text{A C G T A C - - - G T A C G T A C} \]

Read: \[ \text{A C G T A C T A G G T A C G T A C} \]
Example 4: indels

Reference: ACGTAC- - GTACGTAC

Read: ACGTACGTAC
Indel detection: rules

\[ u = vwxy \]
Consequently...

It’s hard to decide the cause of a color mismatch without having the memory of the previous colors in the sequence.
Reads and their qualities are loaded into memory

The reference genome is translated in colors and fully indexed

Seeds are passed on each read (on the established positions, if such positions are mentioned, or on all positions otherwise)

For each hit, a semi-global alignment is performed between the read and the corresponding part of the reference

The best \( M \) such alignments are memorized for each read (\( M = 4 \) by default)

Afterwords, the reads are mapped to the genome using a simple greedy algorithm, and the mapped base sequence is built from this mapping (reading errors are ignored, SNPs and indels are used to determine the new sequence)
Filtering

- Step 1: Seeds designed with Iedera, based on a model adapted to the sequence length and error distribution
- Step 2: For all hits, fast SIMD alignment (Laurent)

If successful, build full alignment, detecting base mutations and reading errors.
Processing a hit

hit

reference genome

read
Processing a hit

Mapping reads from AB Solid sequencers
Alignment matrix

Reference genome fragment

Read
There are two overlapped “models” that cause mismatches:

- Reading errors: one misread color, if interpreted, completely chances the nucleotide sequences
- Actual differences between the sequenced data and the reference: SNPs, insertions, deletions.

The algorithm must explicitly acknowledge and treat these two models, in order to properly distinguish various mismatch types.

This makes the recurrence relations in the dynamic programming algorithm much more complex than the basic Smith-Waterman.
Facts about the reading errors

- A reading error that has appeared on position $i$ is very likely to be followed by other reading errors at positions $i + 5, i + 10, ...$ (due to readings being performed in cycles of 5)

- The chance of errors is usually higher at the end of the read
Facts about mutations

- the chance of having a sequence of $n$ consecutive base mutations decreases exponentially in $n$
Data structures

The information that needs to be memorized in each case of the matrix:

- the score of the partial alignment ending there
- the number of indels on this path so far
- the color resulted by composing all the pairs in the path should be BLUE (identity) if there are no reading errors and the current case is not “in the middle” of a SNP
- the color change at the beginning of the last sequence thought to be a SNP sequence
- the “freshness” of that color change, in number of cases on the path that have been visited since; depends on the read quality
- for quicker traceback, the relative position of the previous case is memorized
The match/mismatch scores depend on the read quality. The lower the quality, the lower the match reward. The lower the quality, the lower the mismatch penalty.

<table>
<thead>
<tr>
<th>Quality</th>
<th>[0..2]</th>
<th>(2..5)</th>
<th>(5..10)</th>
<th>(10..40+]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match score</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Mismatch penalty</td>
<td>0</td>
<td>−1</td>
<td>−2</td>
<td>−3</td>
</tr>
</tbody>
</table>
Algorithm

- Classic initialisation for semi-global alignment
- The computation takes into account the colors preceding the ones being aligned at each step
Example 1: identical sequences

Reference sequence:

```
reference
```

```
read
```

Mapping reads from AB Solid sequencers
**Example 1: identical sequences**

![Sequence Diagram]

- **Reference** sequence (top line)
- **Read** sequence (bottom line)
Example 1: identical sequences

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 1: identical sequences

Reference sequence:

Read sequence:
Example 1: identical sequences

---

**reference**

---

**read**
Example 2: 1 reading error

reference

read
Example 2: 1 reading error
Example 2: 1 reading error

Mapping reads from AB Solid sequencers
Example 2: 1 reading error

Reference: [Diagram showing a sequence with blue, green, and red symbols.]

Read: [Diagram showing a sequence with blue, green, and red symbols.]

Mapping reads from AB Solid sequencers
Example 2: 1 reading error

Mapping reads from AB Solid sequencers
Example 2: 1 reading error

```
reference
---
read
```

Mapping reads from AB Solid sequencers
Example 2: 1 reading error

Mapping reads from AB Solid sequencers
Example 2: 1 reading error

Mapping reads from AB Solid sequencers
Example 2: 1 reading error
Example 3: 2 consecutive mutations
Example 3: 2 consecutive mutations

Reference:

Read:
Example 3: 2 consecutive mutations
Example 3: 2 consecutive mutations
Example 3: 2 consecutive mutations

Mapping reads from AB Solid sequencers
Example 3: 2 consecutive mutations

Reference:

Read:
Example 3: 2 consecutive mutations
Example 4: reading error + SNP

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 4: reading error + SNP

reference

read
Example 4: reading error + SNP

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 4: reading error + SNP

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 4: reading error + SNP

Reference sequence:

Read sequence:

Mapping reads from AB Solid sequencers
Example 4: reading error + SNP

reference

read
Example 4: reading error + SNP

<table>
<thead>
<tr>
<th>reference</th>
<th>read</th>
</tr>
</thead>
<tbody>
<tr>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
<tr>
<td>🟠</td>
<td>🟠</td>
</tr>
</tbody>
</table>

Mapping reads from AB Solid sequencers
Example 4: reading error + SNP

Reference sequence:

Read sequence:
Example 4: reading error + SNP
Example 5: deletion

Mapping reads from AB Solid sequencers
Example 5: deletion
Example 5: deletion

Reference:

Read:
Example 5: deletion

Mapping reads from AB Solid sequencers
Example 5: deletion

Mapping reads from AB Solid sequencers
Example 5: deletion

Mapping reads from AB Solid sequencers
Example 5: deletion

Mapping reads from AB Solid sequencers
Example 6: reading error + 1 deletion

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 6: Reading error + 1 deletion

Mapping reads from AB Solid sequencers
Example 6: reading error + 1 deletion

reference:

read:
Example 6: reading error + 1 deletion

reference

read
Example 6: reading error + 1 deletion

Reference:

Read:

Mapping reads from AB Solid sequencers
Example 6: reading error + 1 deletion

[Diagram showing a reference sequence with one deletion and a corresponding read sequence]

Mapping reads from AB Solid sequencers
Example 6: reading error + 1 deletion

### Reference
- Blue: Match
- Green: Insertion
- Red: Deletion
- Yellow: Substitution

### Read
- Blue: Match
- Green: Insertion
- Red: Deletion
- Yellow: Substitution
Example 6: reading error + 1 deletion
Example 6: reading error + 1 deletion

Mapping reads from AB Solid sequencers
Perspectives

- Improve detection of overlapping SNPs and reading errors
- Improve mapping step: find the best mapping of a read based on the reads already mapped in that zone