Detection of viral integration sites

Corinna Blasse

Advanced Algorithms for Bioinformatics (P4)
Instructor: Birte Kehr

SoSe 2010
FU Berlin

July 18, 2010
Biological situation

- **Human Herpesvirus Six** (HHV-6) is a wide spread virus
- One possible persistence by virus integration

Virus integration: Insertion of viral DNA into host-cell DNA
Experiment

- Integration of human herpesvirus 6 into the human genome
- Sequencing with the SOLID 3 Plus System

**Goal:** To find reads containing virus and host genome
Outline of the project

Input:
- Millions of SOLID reads
- Sequence of the host genome
- Sequence of the viral genome

Steps of the project:
- Develop a local read mapper for color-space reads
- Apply the read mapper to find viral integration sites in the human genome

Output:
Color-space reads

= Reads that contain one of four colors for every pair of nucleotides

\[ T0001201030 \]

Encoding by using the AB SOLID sequencing system

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Diagram showing the encoding system with nodes A, C, G, T and arrows indicating the transitions between colors.
Local Read Mapper for color-space reads

1. Import the color-space reads and convert them
2. Import the sequences of the virus and human genome
3. Read mapping by using Swift-Local in SeqAn
4. Filter the mapped reads
   - Complete mapped reads
   - Local mapped reads
5. Construct reads that belong to both genomes
Results

- Mapping of a test data set containing reads of the following structures
  - Human genome
  - Virus genome
  - Virus genome | Human Genome
  - Human Genome | Virus Genome

- Just handling of sequence errors, not errors in the color-code
Sources

*Principles of Di-Base Sequencing and the Advantages of Color Space Analysis in the SOLiDTM System.*  

Rumble SM & Lacroute P & Dalca AV & Fiume M & Sidow A et.al.  
*SHRiMP: Accurate Mapping of Short Color-space Reads.*  