VARiD: Variation Detection in Color-Space and Letter-Space

Adrian Dalca¹ and Michael Brudno¹,²
University of Toronto

¹ Department of Computer Science
² Banting & Best Department of Medical Research
Motivation

we have different Color-space and Letter-space platforms

Methods

Results

Advantages
Sequencing Platforms

- **letter-space**
  - Sanger, 454, Illumina, etc
  - not as many software tools out there

- **color-space**
  - AB SOLiD
  - different sequencing biases, different inherent errors and different advantages
  - useful to combine this information
motivation | methods | results | advantages

color space and letter space platforms

**Color Space**

<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>
Color Space

Translating

T2123132303132321213111120
TCAGCATCGGCATCGACTGCACAGG

Sequencing Error vs SNP

T2123132303132321213111120
T2123132303102321213111120
TCAGCATCGGCAAGCGACTGCACAGG
TCAGCATCGGCAAGCTGACGTGTCC
Notes:
• clear distinction between a sequencing error and a SNP
• can this help us in SNP detection? sounds like it!
  single color change → error,
  2 colors changed → (likely) SNP.

Example

**TTTTT GAGAGGAATA**

**TTTTT GAGAGGAATA**

**Sequencing Errors**

**SNP**

**VARiD toolbox GUI**
Examples (more realistically)

reference

A C T

guess: Het SNP

reads

real data

e.g. above

heterozygous SNPs

a lot more errors
Motivation

- we want a SNP caller to handle both traditional letter-space as well as color-space reads

Realistically, situation is tougher.

- Heterozygous SNPs
- Homologous SNPs
- Tri-allelic SNPs
- small indels
- alot more error than in original previous example
- misalignment (by chance)
- misalignment (consistently)
Motivation

Methods
Model the system with an HMM
Expand the HMM and apply Heuristics

Results

Advantages

Quick breath.
Hidden Markov Model

Statistical model for a system (so we have states)
Assume that system is a Markov process with state unobserved.
Markov Process: future state depends only on current state
We can observe the state’s emission (output)
each state has a probability distribution over outputs

apply: we don’t know the state (donor?),
but we can observe some output
determined by the state (reads?)
Our Hidden Markov Model

At every pair of consecutive positions:
- don’t know the donor nucleotides,
- have some color-space and/or letter-space reads

The donor could be:
- letters: AA color 0
- letters: AC color 1
- letters: TT color 0

16 combinations

Note: AA and TT give the same colors! So we have redundancy.
Colors and Letters

- AA and TT give the same colors! So we have **redundancy**.
- Can’t just call colors, since they can represent one of several translations.
- To properly call SNPs, we need to **model underlying letters**.
Consider donor at positions 532, 533 and 534. At each pair we have one color, two letters.

16 states

only certain transitions allowed

each state depends on the previous states, but not further (Markov Process)
Unknown genome: ...NNNNNNNNNNNNNNN...

Color reads:
- T0102010 311223
- T103010 311223
- T2010 311223

Color emissions:

Letter reads:
- ATTGC
- TGGC
- GCGCA

Letter emissions:

Methods and heuristics: HMM models and heuristics
Our Hidden Markov Model

### Emissions

<table>
<thead>
<tr>
<th>Emission</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>color 0</td>
<td>$1 - \frac{\varepsilon}{3}$</td>
</tr>
<tr>
<td>color 1</td>
<td>$\varepsilon$</td>
</tr>
<tr>
<td>color 2</td>
<td>$\varepsilon$</td>
</tr>
<tr>
<td>color 3</td>
<td>$\varepsilon$</td>
</tr>
<tr>
<td>letters A</td>
<td>$(1 - \frac{\xi}{3})$</td>
</tr>
<tr>
<td>letters C</td>
<td>$\xi$</td>
</tr>
<tr>
<td>letters G</td>
<td>$\xi$</td>
</tr>
<tr>
<td>letters T</td>
<td>$\xi$</td>
</tr>
</tbody>
</table>

- **Same distribution of emissions in color-space**
- **Different emissions in letter-space**
Emissions Probability

How do we use emissions?
Assign an **Emission Probability** to each state:
What is the probability that this state emitted these reads.

E.g. For state CC:

\[
p_E = \left(1 - \frac{\varepsilon}{3}\right)^2 \times \varepsilon^1 \times \left(1 - \frac{\xi}{4}\right)^1 \times \xi^2
\]
So we have
• the **unknown** (donor pair at some location),
• the **emissions** (output – the read colors at some location), and
• the dependency on the **previous state**.
Our Hidden Markov Model

- Have set-up a form of an HMM
- run Forward-Backward algorithm
- get probability distribution over states

likely state
Current form of HMM only detects homozygous SNPs

We include:
- short indels
- **heterozygous** SNPs
Expansion: Gaps and heterozygous SNPs

Expand states
- Have states that include gaps
  - emit: gap or color
- Have larger states, for diploids
  - emit: colors

Same algorithm, but in all we have 1600 states
Expansion: Gaps and heterozygous SNPs

• Use variable error rates for emissions
  o can support quality values (alter the emission probabilities)

• Translate through the first letter
  o gives guidance in letter-space
  o know the error rate (= error rate at first color)
  note: not ok to translate the whole read due to effects of color-space error, but one letter is safe. handle like a normal letter-space emission

```
> T2 12313230312332121311120
>> C 12313230312332121311120
```
Post Processing: Uncorrelated Errors

HMM doesn’t know which read each emission came from.

Example

We will get a lot of confidence in states voting for which is a het SNP

But there are NO reads supporting Blue-Green

Post Processing: For each proposed variant, check that there actually is enough reads supporting this variant. Several other cases are handled with a similar check.
Results

Quicker breath.
Working Results

Simulations

Color-space dataset
- **Source**: JCVI. Validated with Sanger. Mappings are done with SHRiMP
- 8 datasets all with similar performance:
  - 83-87% True Positives (real SNPs called)
  - few False Positives (non-var called as SNPS) --- 10-15% of calls, 0.02% of nucleotides
  - results very similar to Corona;

Examples (~25000 bp)

<table>
<thead>
<tr>
<th></th>
<th>NA19137</th>
<th></th>
<th>NA18504</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>VARiD</td>
<td>38/44</td>
<td>10</td>
<td>54/65</td>
<td>7</td>
</tr>
<tr>
<td>Corona</td>
<td>39/44</td>
<td>10</td>
<td>55/65</td>
<td>10</td>
</tr>
</tbody>
</table>
motivation | methods | results | advantages

Example of False Positive

Sanger (“real”) haplomes

Color-space Reads

VARiD Het SNP Prediction

Example of False Negative (missed call)

Sanger (“real”) haplomes

Color-space Reads

VARiD Prediction

Example of False Positive

Sanger (“real”) haplomes

Color-space Reads

VARiD Het SNP Prediction

Example of False Negative (missed call)

Sanger (“real”) haplomes

Color-space Reads

VARiD Prediction
Advantages

take advantage of both Color-space and Letter-space reads

Adjacent SNPs, short indels

Quicker breath.
Summary of VARiD

• Treats color-space and letter-space together in the same framework
  • no translation – take advantage of each technology’s properties
  • fully probabilistic

• Handles adjacent SNPs

Example
    reference  CAAAG translates to C1 02
    donor      CTTTG translates to C2 01

Looks like 2 sequencing errors.
VARiD can detect the 2 SNPs
Find us @ the poster session: U61.
Monday (June 29) evening

VARiD website
http://compbio.cs.utoronto.ca/varid

Thank you:
Sam Levy at JCVI
NSERC

Contact:
dalca@cs.utoronto.ca