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

Adrian Dalca[1] and Michael Brudno[1,2]

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

need oth)

Motivation

Methods

Results

Advantages

combining the different Color-space and Letter-space platforms

Sequencing Platforms

- letter-space
  Sanger, 454, Illumina, etc

```
> NC_005109.2 | BRCA1 SX3
TCAGCATCGGCATCGACTGCACAGG
```

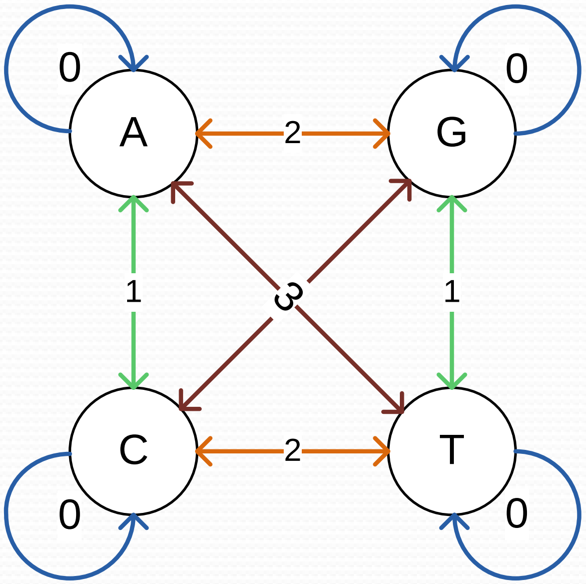
- color-space
  AB SOLiD
  not as many software tools out there

```
> NC_005109.2 | BRCA1 AF3
T212313230313232121311120
```

- different sequencing biases, different inherent errors and different advantages
  - useful to combine this information

Color Space

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0



pic reference: SHRiMP

Color Space

Translating

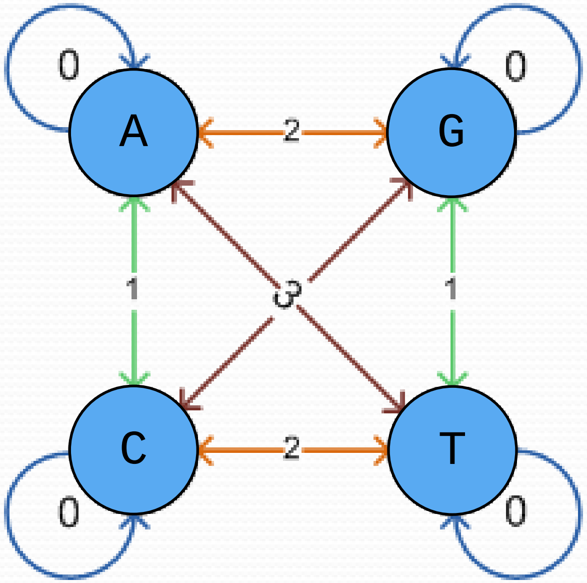
T212313230313232121311120
TCAGCATCGGCATCGACTGCACAGG

Sequencing Error vs SNP

> T212313230313232121311120
> T21231323031**0**232121311120

> TCAGCATCGGCA**G**CGACTGCACAGG
> T21231323031**23**32121311120

> T21231323031**0**232121311120
> TCAGCATCGGCA**AGCTGACGTGTCC**



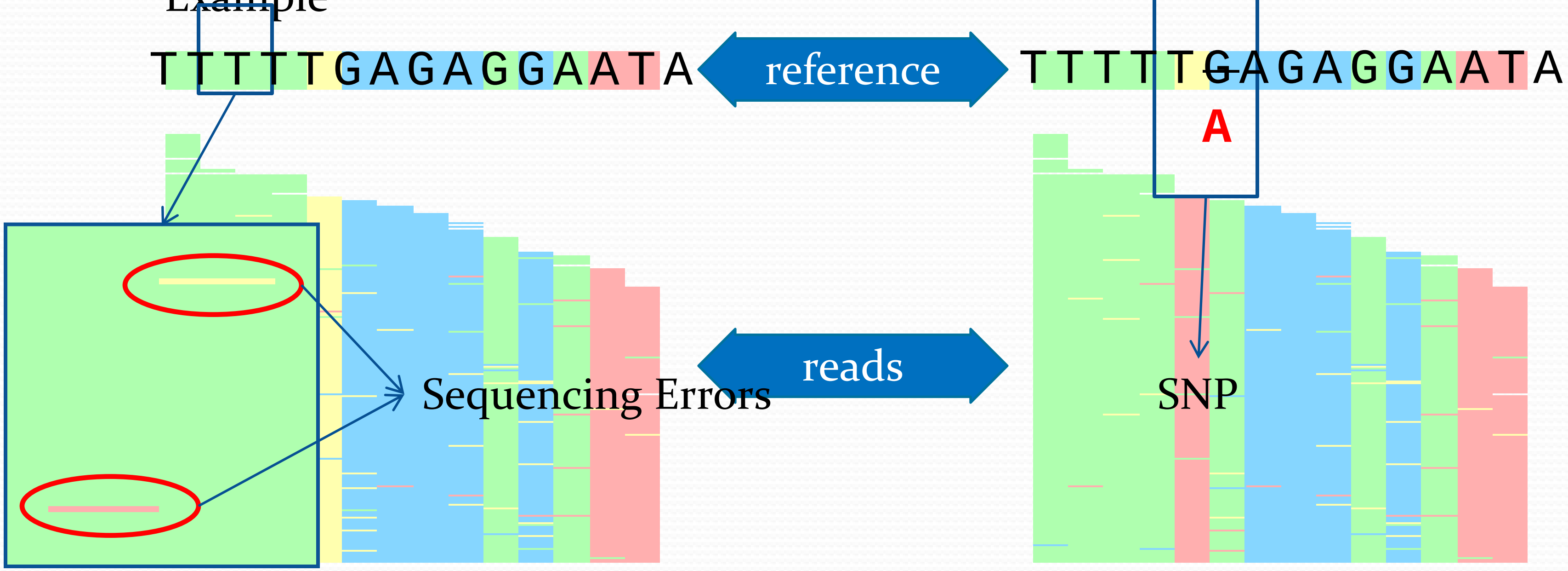
Notes:

- clear distinction between a sequencing error and a SNP
- can this help us in SNP detection? sounds like it!

  single color change → error,

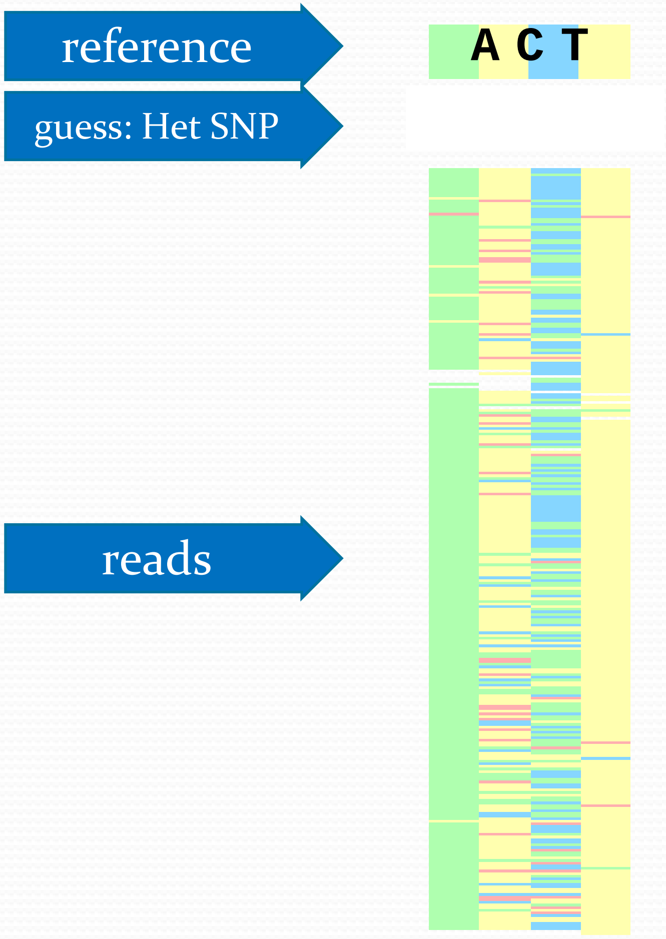
  2 colors changed → (likely) SNP.

Example



VARiD toolbox GUI

Examples (more realistically)

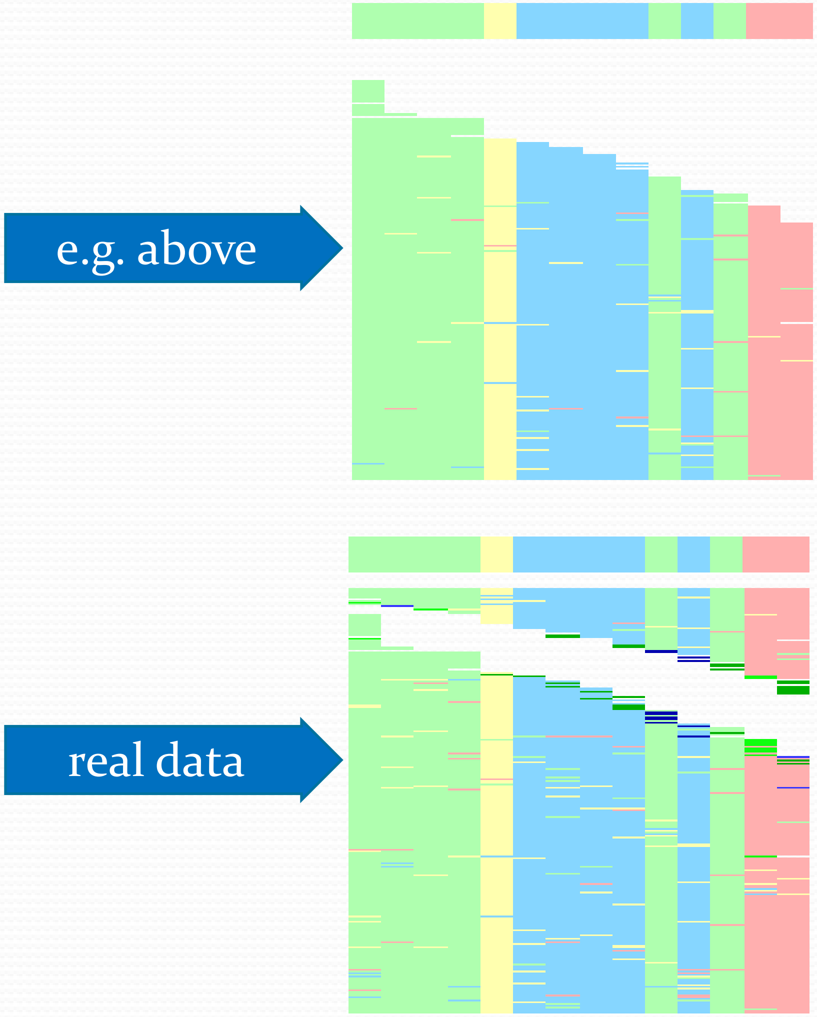


reference

guess: Het SNP

reads

heterozygous SNPs



e.g. above

real data

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

(for colors)

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

  :

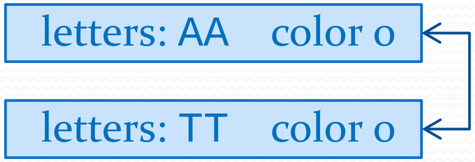
- letters: TT   color 0

16 combinations

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

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

Colors and Letters

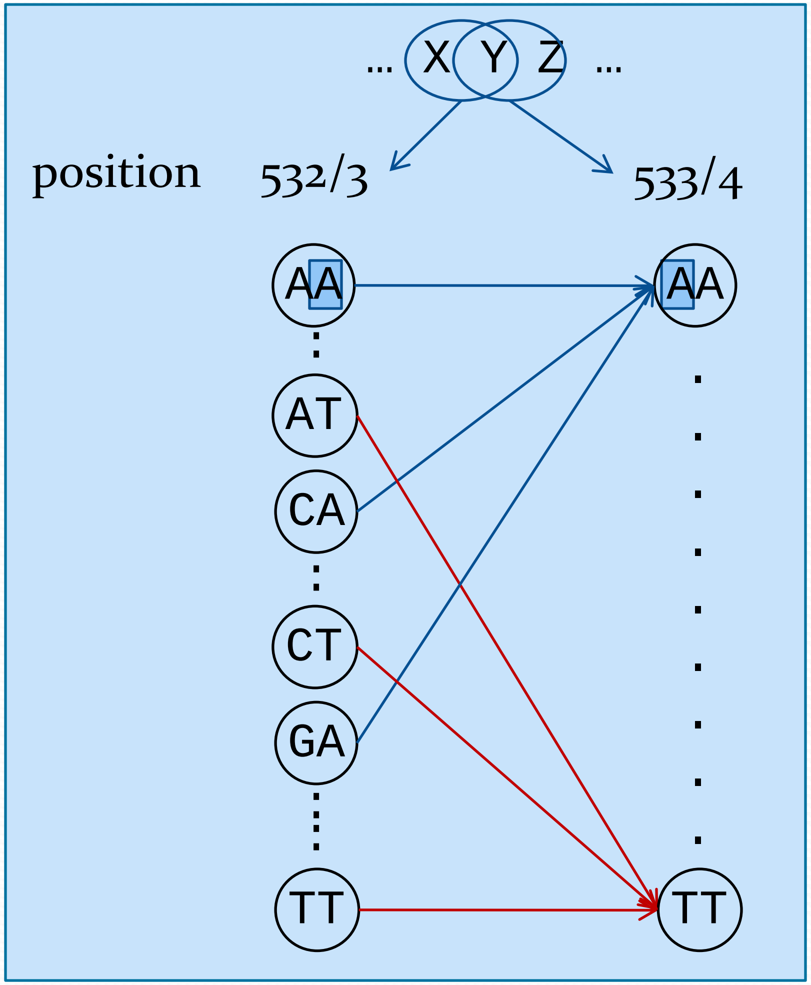


AA and TT give the same colors!
So we have redundancy.

- can’t just call colors, since they can represent one of several translations

- to properly call SNPs, we need to model underlying letters.

States of the Model



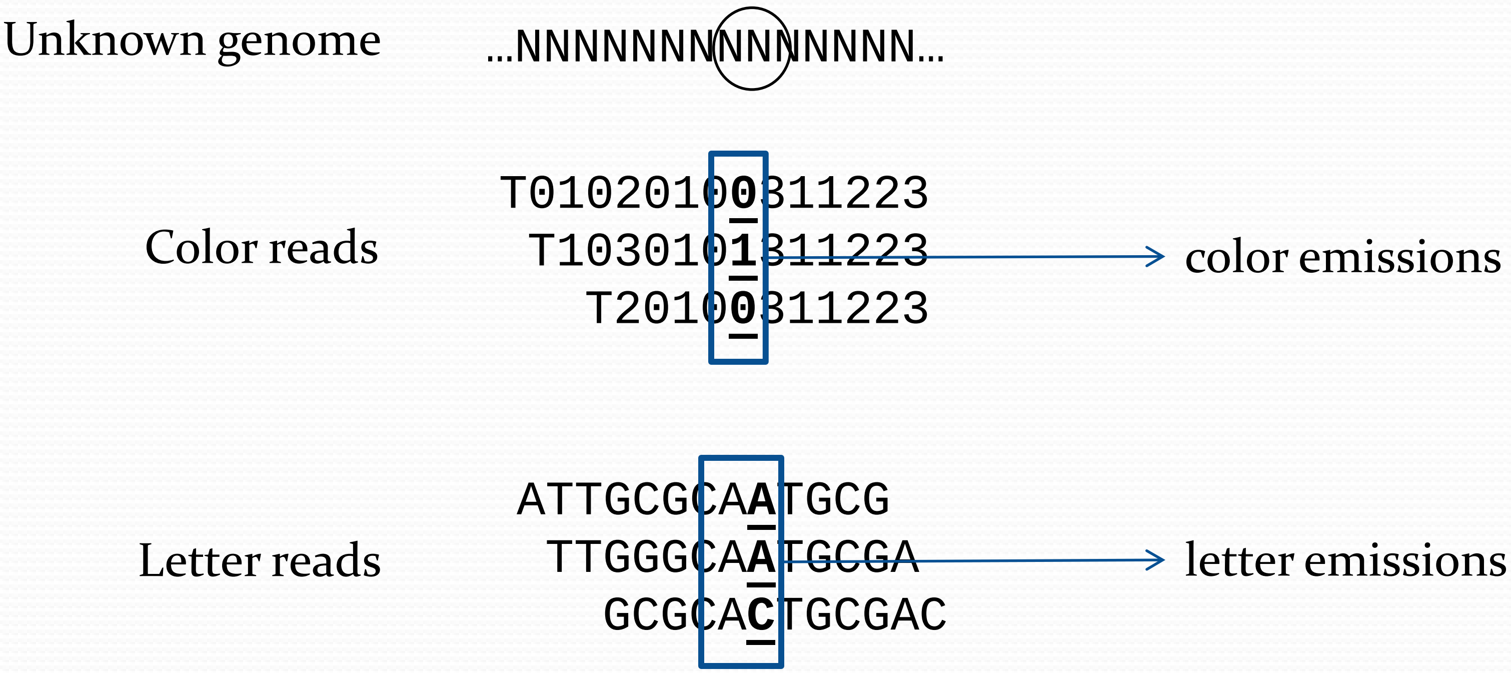
Consider donor at positions 532, 533 and 534.
At each pair we have one color, two letters

16 states

only certain transitions allowed

each state depends on the previous states, but not further (Markov Process)

Emissions



Unknown genome ...NNNNNNNNNNNNNN...

Color reads

T0102010**0**311223
T103010**1**311223
T2010**0**311223

color emissions

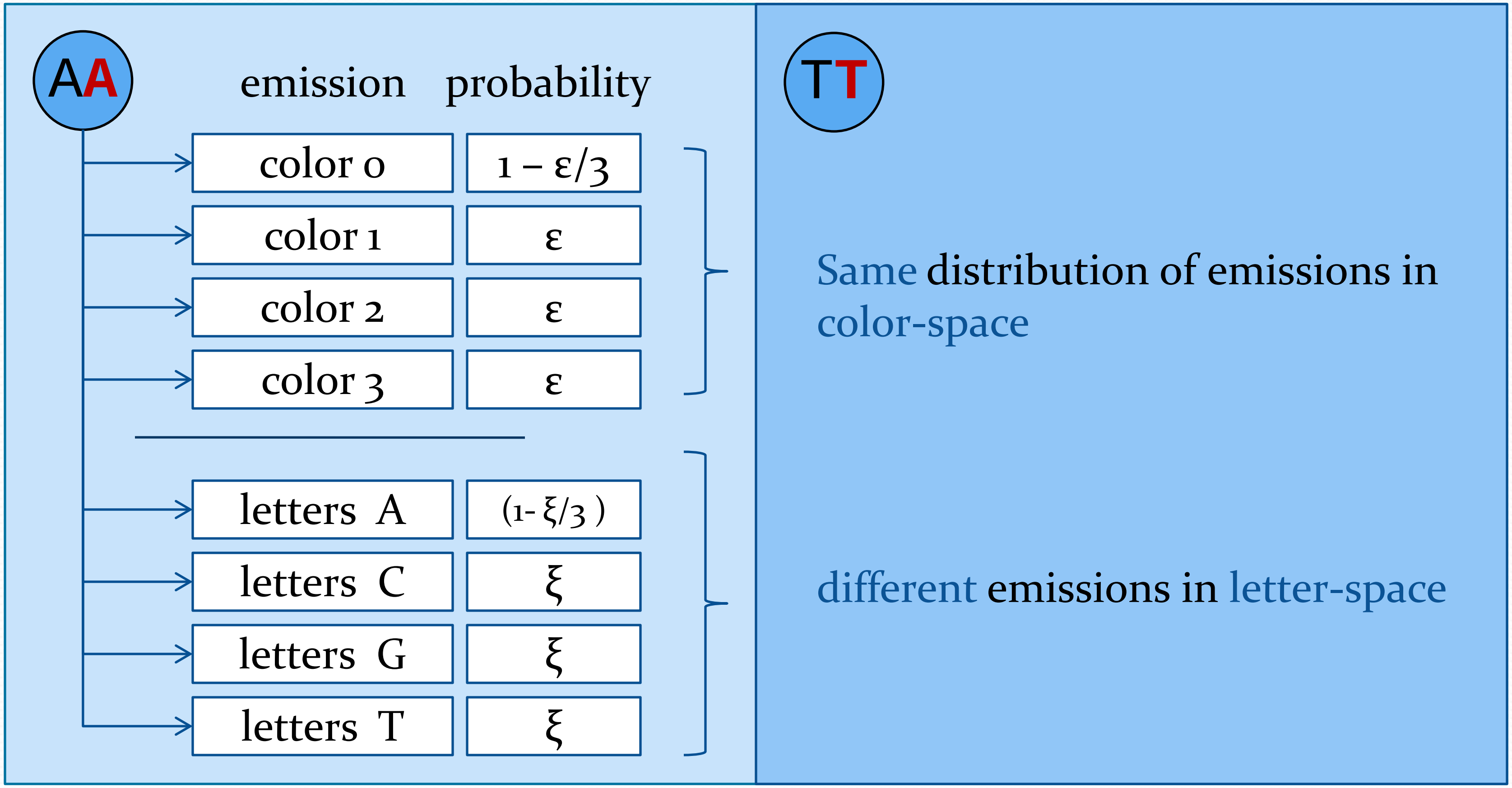
Letter reads

ATTGCGCA**A**TGCG
TTGGGCA**A**TGCGA
GCGCA**C**TGCGAC

letter emissions

Our Hidden Markov Model

Emissions



AA

emission	probability
color 0	1 – ε/3
color 1	ε
color 2	ε
color 3	ε
letters A	(1- ξ/3)
letters C	ξ
letters G	ξ
letters T	ξ

TT

Same distribution of emissions in color-space

different emissions in letter-space

Emissions Probability

```
...NNNNNNNNNNNNNN...

T01020100311223
 T10301013112223
   T20100311223
```

How do we use emissions?

Assign an Emission Probability to each state:

What is the probability that this state emitted these reads.

```
ATTGCGCAATGCG
 TTGGGCAATGCGA
   GCGCACTGCGAC
```

E.g. For state CC:

$$p_E = [(1-\frac{\varepsilon}{3})^2 \times \varepsilon^1] \times [(1-\frac{\xi}{4})^1 \times \xi^2]$$

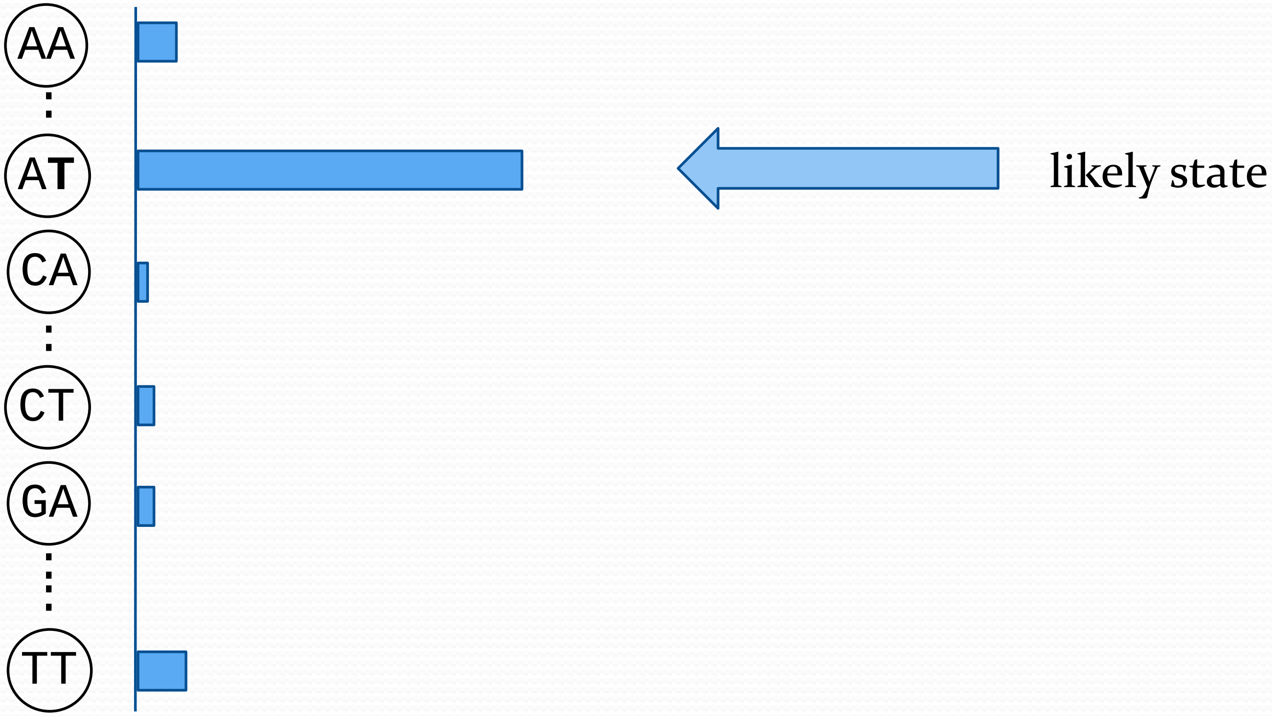
Our Hidden Markov Model

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



Current form of HMM only detects homozygous SNPs

We include :

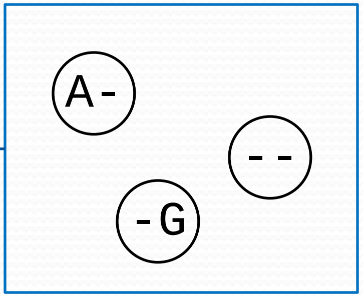
- short indels
- heterozygous SNPs

Expansion and Heuristics

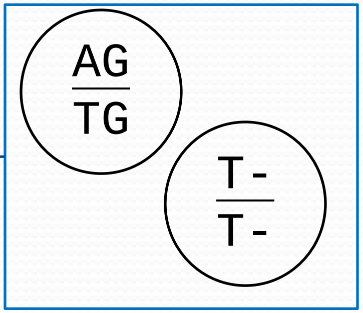
Expansion: Gaps and heterozygous SNPs

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
  - can support quality values (alter the emission probabilities)

- Translate through the first letter
  - gives guidance in letter-space
  - know the error rate (= error rate at first color)

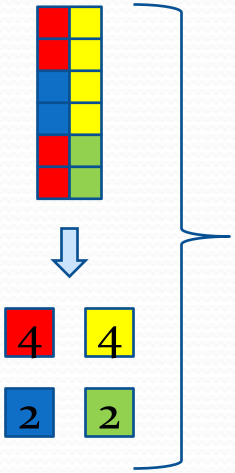
    note: not ok to translate the whole read due to
    effects of color-space error, but one letter is safe.
    handle like a normal letter-space emission

```
>T212313230312332121311120
>>C12313230312332121311120
```

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.

Motivation

Methods

Results

Advantages

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)

	NA19137		NA18504	
	TP	FP	TP	FP
VARiD	38/44	10	54/65	7
Corona	39/44	10	55/65	10

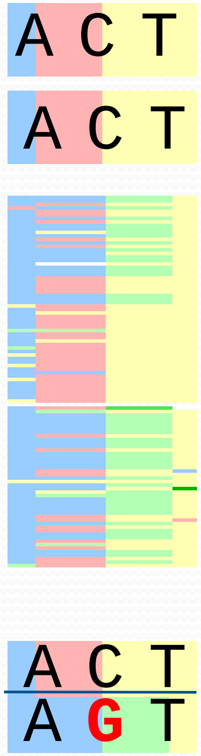
simulations and real data

Example of False Positive

Sanger ("real") haplomes

Color-space Reads

VARiD Het SNP Prediction

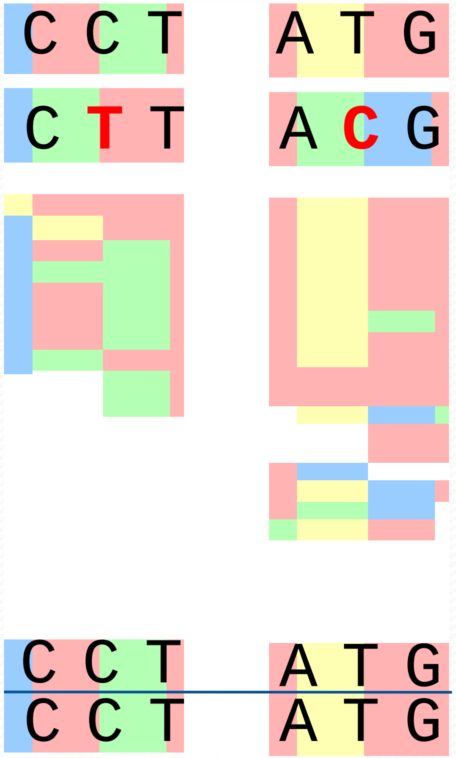


Example of False Negative (missed call)

Sanger ("real") haplomes

Color-space Reads

VARiD Prediction



Motivation

Methods

Results

Advantages

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

Adjacent SNPs, short indels

Quicker breath.

combining both platforms natively

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 **CAAG** translates to **C102**

donor **CTTG** translates to **C201**

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

VARiD

Adrian Dalca & Michael Brudno
University of Toronto