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

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

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

University of Toronto

In this poster, we present VARiD - a Hidden Markov Model for SNP and indel identification with AB-SOLiD color-space as well as regular letter-space reads. VARiD combines both types of data in a single framework which allows for accurate predictions.

Motivation

There are two types of sequencing methodologies: letter-space (Sanger, 454, Illumina, etc) and color-space (AB SOLiD). They have different sequencing biases, different inherent errors and different advantages, and **we combine information from these platforms**.

```
> letter_space_eg
TCAGCATCGGCAT
> color_space_eg
T212313230313
```

Color Space Properties

In color-space, a color is given for each pair of base-pairs (bp). There are 4 colors for 16 bp combinations, as shown by the matrix to the right. For example, an A followed by a G is represented by the color 2. Certain properties arise:

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

- a sequencing error is a single color change

```
> T212313230313232121311120
> T212313230310232121311120
```

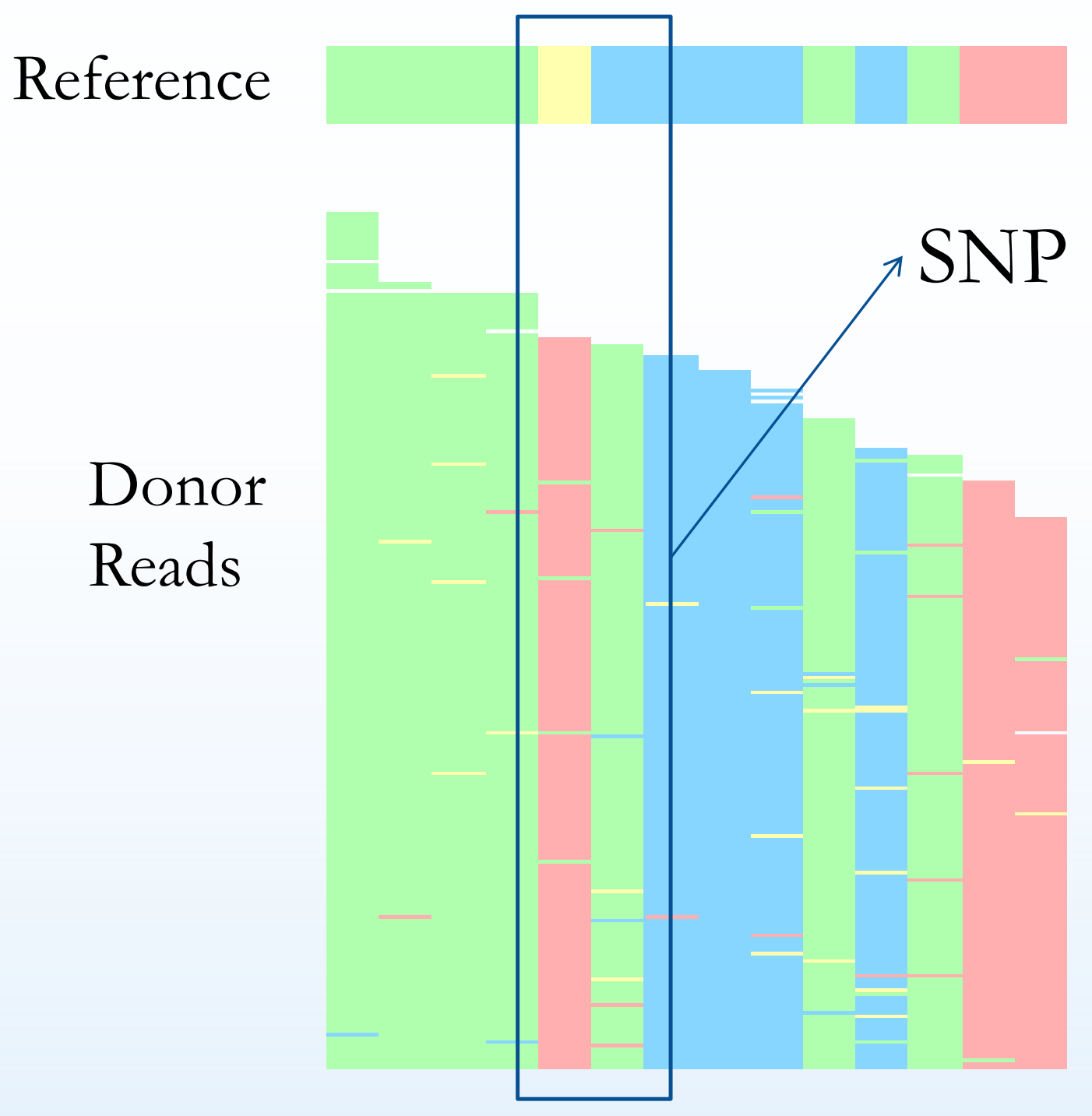
- a SNP represents two color changes

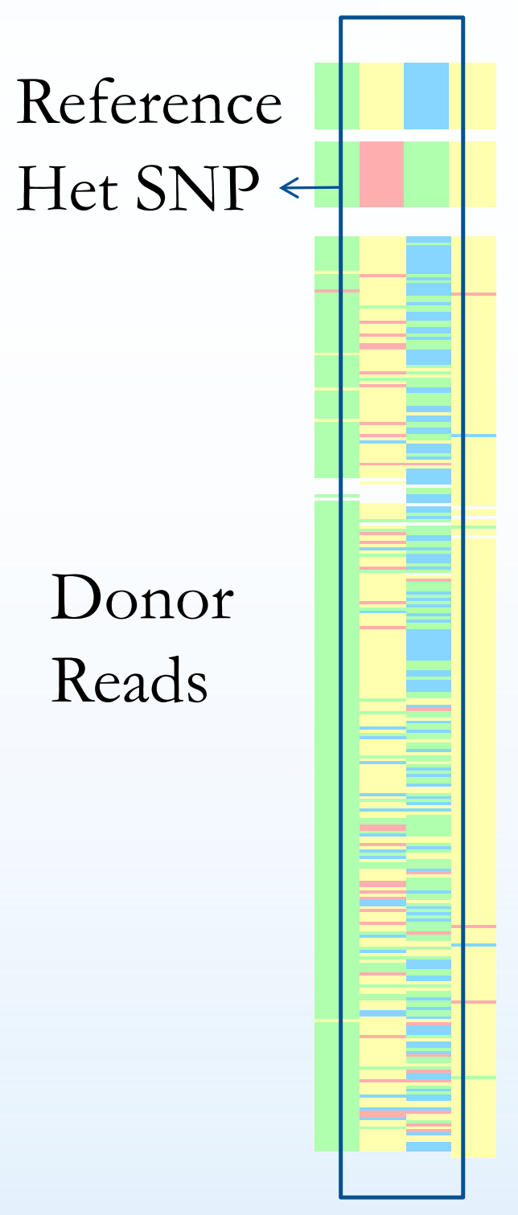
```
> TCAGCATCGGCAGCGACTGCACAGG
> T212313230312332121311120
```

- if we translate a color-space read we get the entire sequence wrong after an error

```
> T212313230310232121311120
> TCAGCATCGGCAAGCTGACGTGTCC
```

These properties may allow us to call SNPs in clear cases. Below we give examples with color-space reference and reads. In the first example the donor reads give a strong, clear signal. The more realistic second example shows a more complicated situation.





Results

For now, we ran VARiD on a color-space datasets from JCVI, with Sanger validation. All of the datasets resulted in similar performance of 83-87% True Positives (real SNPs called) and few False Positives (non-var called as SNPS) i.e. around 10-15% of calls, 0.02% of nucleotides. We note that the results were very similar to running the Corona Lite pipeline, a software from AB SOLiD specifically for color-space reads. Upon manual inspection, many of the missed calls (by either software) are under low or inaccurate coverage.

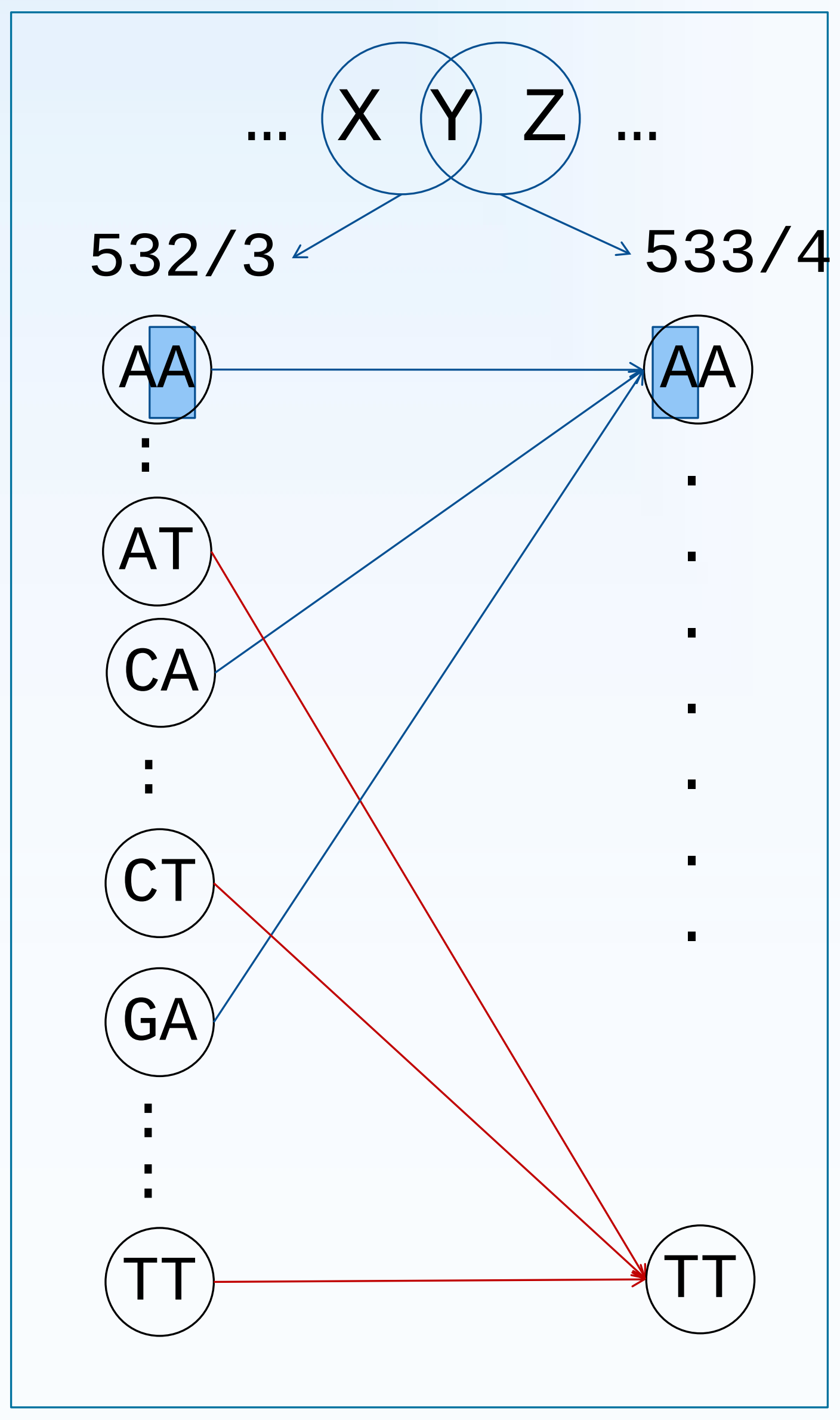
Example results

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

Methods

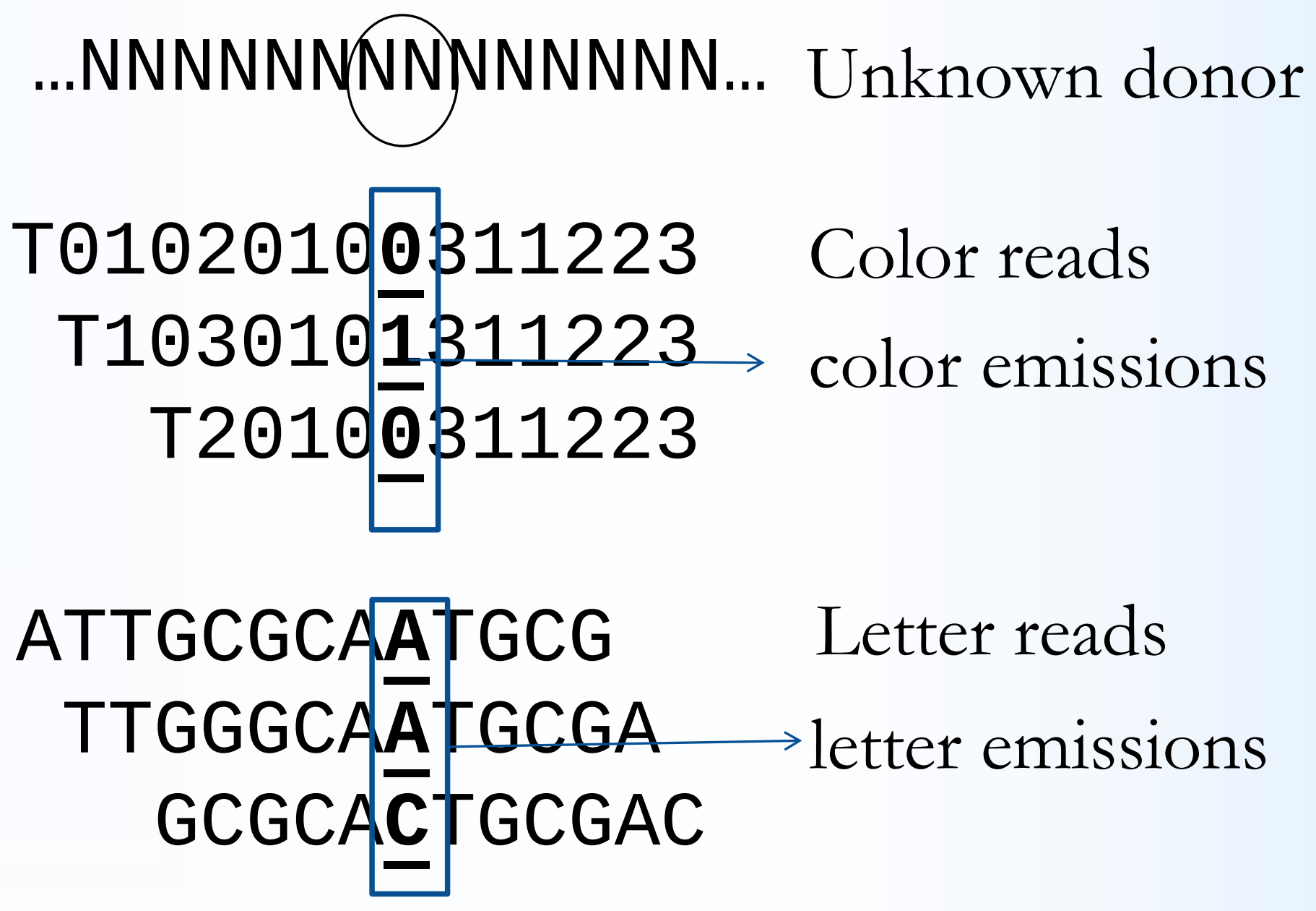
A **Hidden Markov Model** (HMM) is a statistical model for a system (which can be in one of various states and can evolve). We assume that the system is a Markov Process (where a future state depends only on the current state). We cannot see the states directly (they are hidden), but we can observe their emission (output).

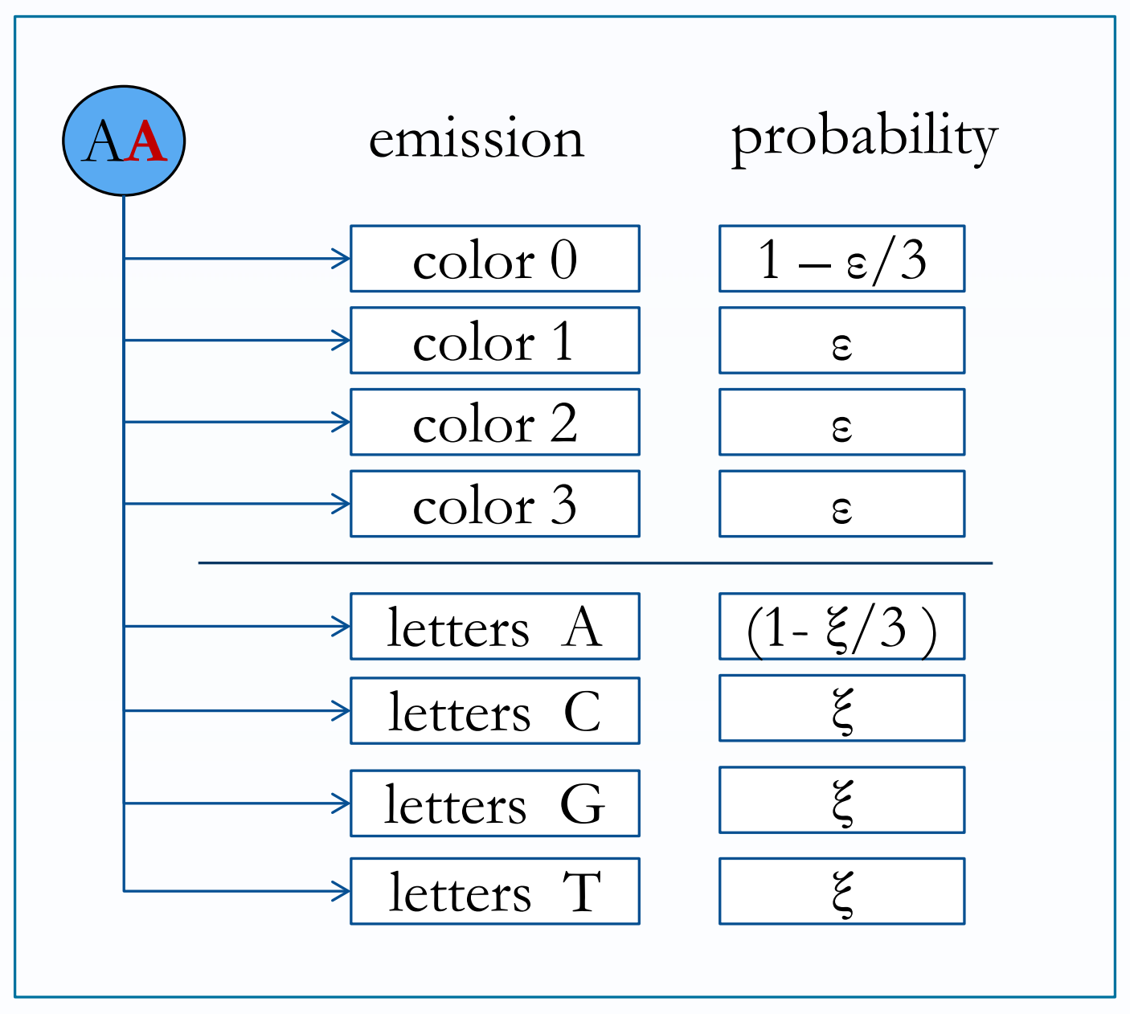
We apply an HMM to our problem: we don't know the donor at a position (unknown state), but we observe reads from the donor (state's emission). We detail a model for the underlying letters.



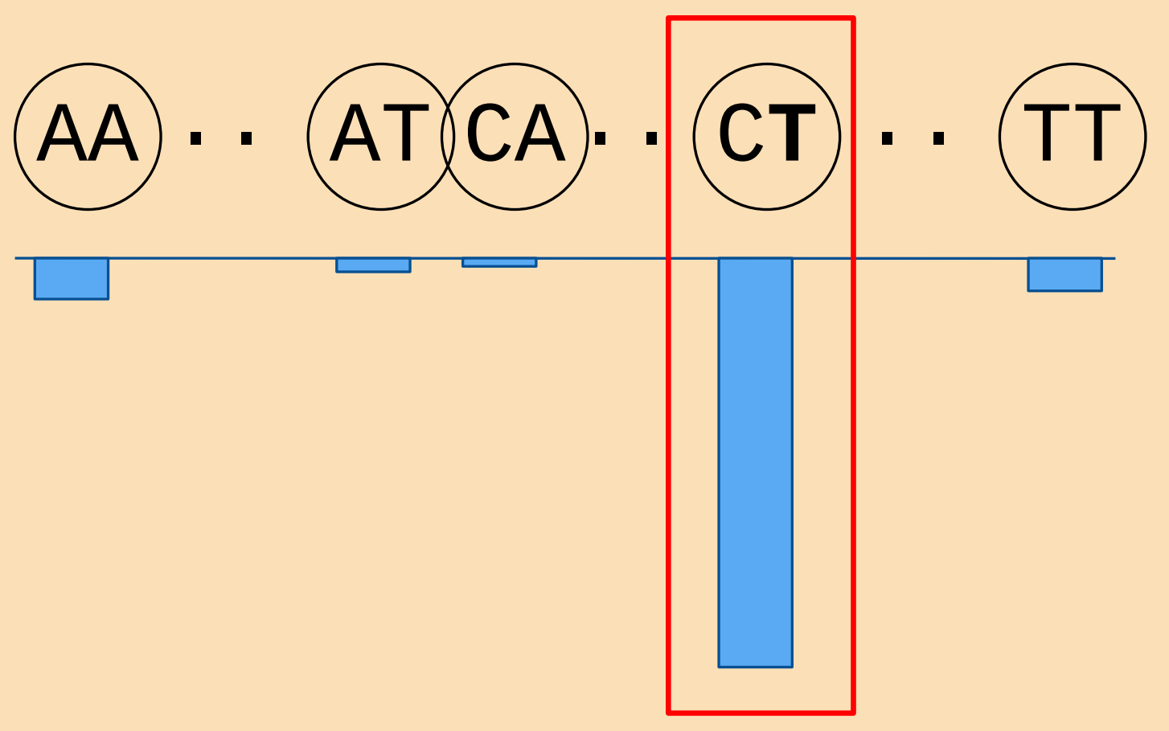
Consider 3 donor positions 532 (X), 533 (Y), and 534 (Z). Nucleotides XY can be any of AA, AC, …TT, and similarly for YZ. Since Y is shared, we can only transition between a state that ends with the same letter that the next state starts with. For example, from state A**A**, we can only transition to a state that start with an A. We note that this is a Markov Process: each state depends only on the previous one.

For an unknown donor, we get emissions via reads: colors from color-space reads, and letters from letter-space reads. (N.B. we overlap pairs - therefore we only need one-letter emissions per pair)





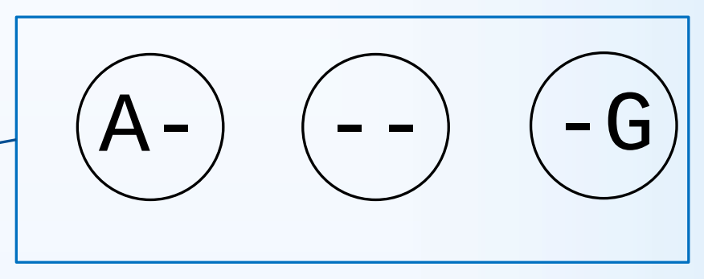
On the left, we see the possible emissions of a state like AA, and the probability that such a state would be emitted. For example, the AA state is very likely to emit the color 0 or letter A, and would only output anything else due to errors.

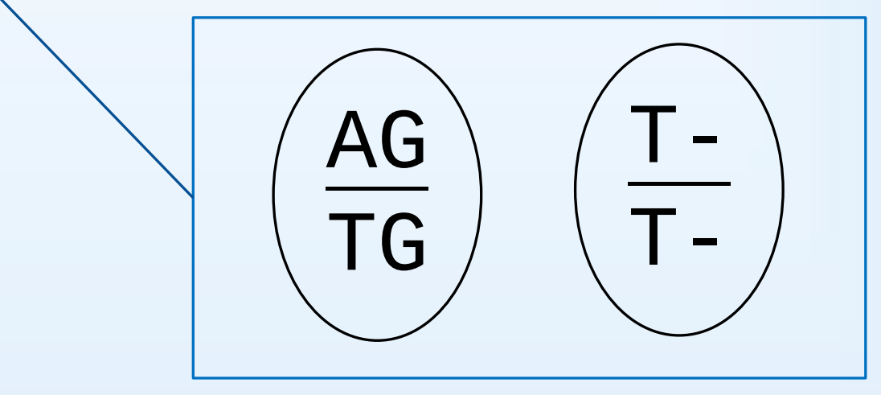


In summary, we have the unknown (donor pair at some location), the emissions (the read colors or letters), and the dependency on the previous state (transitions). We run the Forward-Backward algorithm and get a probability distribution over the possible states.
In the example to the left, we would call the nucleotide at this position T and compare to the reference to determine if it is a SNP.

Next, we **expanded** VARiD to support the following operations:

- to call **small indels**, we add states that can include gaps
- to call **heterozygous SNPs**, we double the size of a state to include two alleles.
- can include a distribution of error rates (and hence quality values)
- we translate through the first color of any color-space read to have letter support in the model





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