VARiD: A Variation Detection Framework for Color- and Letter-Space platforms

Adrian Dalca[1,2], Steven Rumble[3], Sam Levy[4], Michael Brudno[2,5]

1. Massachusetts Institute of Technology
2. University of Toronto
3. Stanford University
4. The Scripps Research Institute
5. Donnelly Centre and the Banting and Best Department of Medical Research

Variation detection from NGS reads

- Determine differences (variation) between **reference** and **donor** using NGS reads of the donor

```
Reference: TCAGCATCGGCATCGACTGCACAGGACCAGTCGATCGAC

Donor:     ???????????????????????????????????????
                    GCATCGACTGCA
                  CGGGATCGACTG
Aligned reads:        ATCCATTGCA
                     GATCCACTGCAC
```

Motivation

Color-space and **Letter-space** platforms
bring them **together**

Methods

Results

Summary

Sequencing Platforms

- letter-space
  Sanger, 454, Illumina, etc

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

- color-space
  AB SOLiD
  less software tools available

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

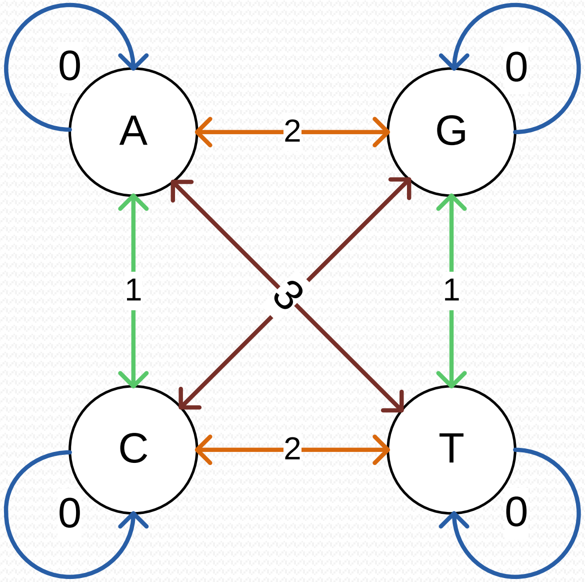
- many differences -> useful to combine this information
  - sequencing biases
  - inherent errors
  - advantages

Color Space

Translation Matrix

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

Translation Automata



```
> T212313230313232121311120
```

Color Space

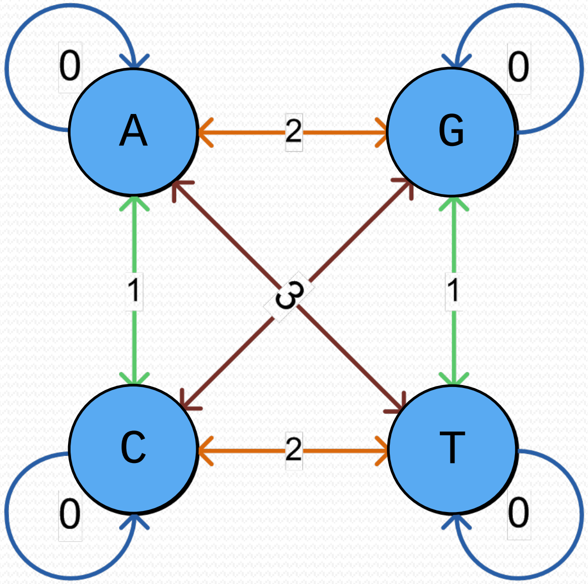
Translating

```
> T212313230313232121311120
> TCAGCATCGGCATCGACTGCACAGG
```

Sequencing Error vs SNP

```
Sequencing Error
> T212313230313232121311120
> T212313230310232121311120
> TCAGCATCGGCAAGCTGACGTGTCC

SNP
> TCAGCATCGGCATCGACTGCACAGG
> TCAGCATCGGCAGCGACTGCACAGG
> T212313230312332121311120
```

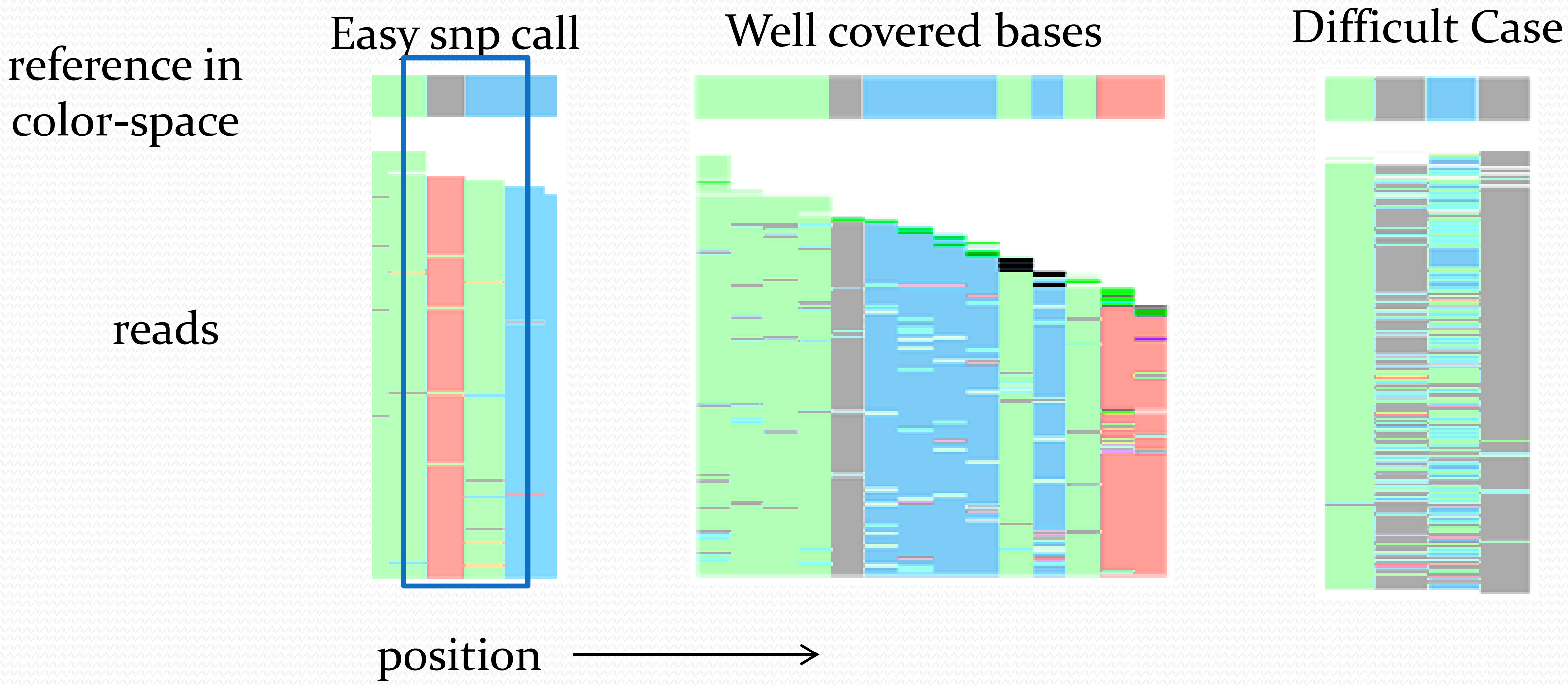


Color Space

- 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.



reference in color-space

reads

Easy snp call

Well covered bases

Difficult Case

position

Motivation

Motivation

- variation caller to handle both letter-space & color-space reads

Detection

- Heterozygous SNPs
- Homozygous SNPs
- Tri-allelic SNPs
- small indels
- account for various errors, quality values & misalignments

VARiD

- system to make inferences on the donor bases
    - variation detection

Motivation

Methods

Simple HMM Model

states, emissions, transitions, FB

Extended HMM Model

gaps, diploids, exceptions

Results

Summary

Hidden Markov Model (HMM)

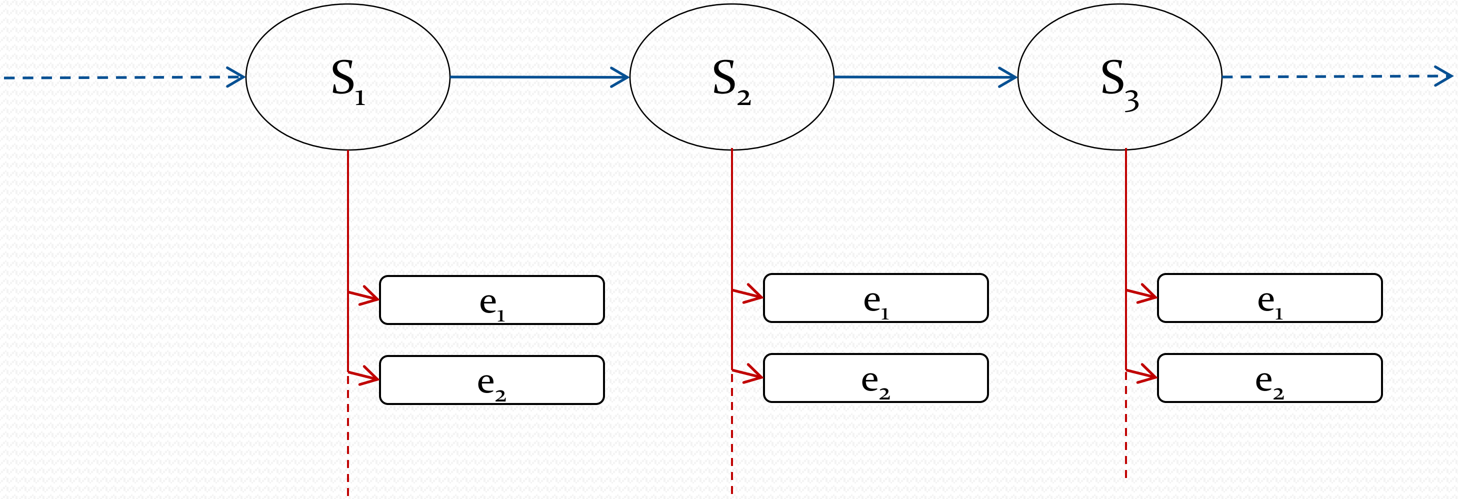
Statistical model for a system - **states**

Assume that system is a **Markov process** with state unobserved.

Markov Process: next state depends only on current state

We can **observe** the state's emission (output)

each state has a probability distribution over outputs

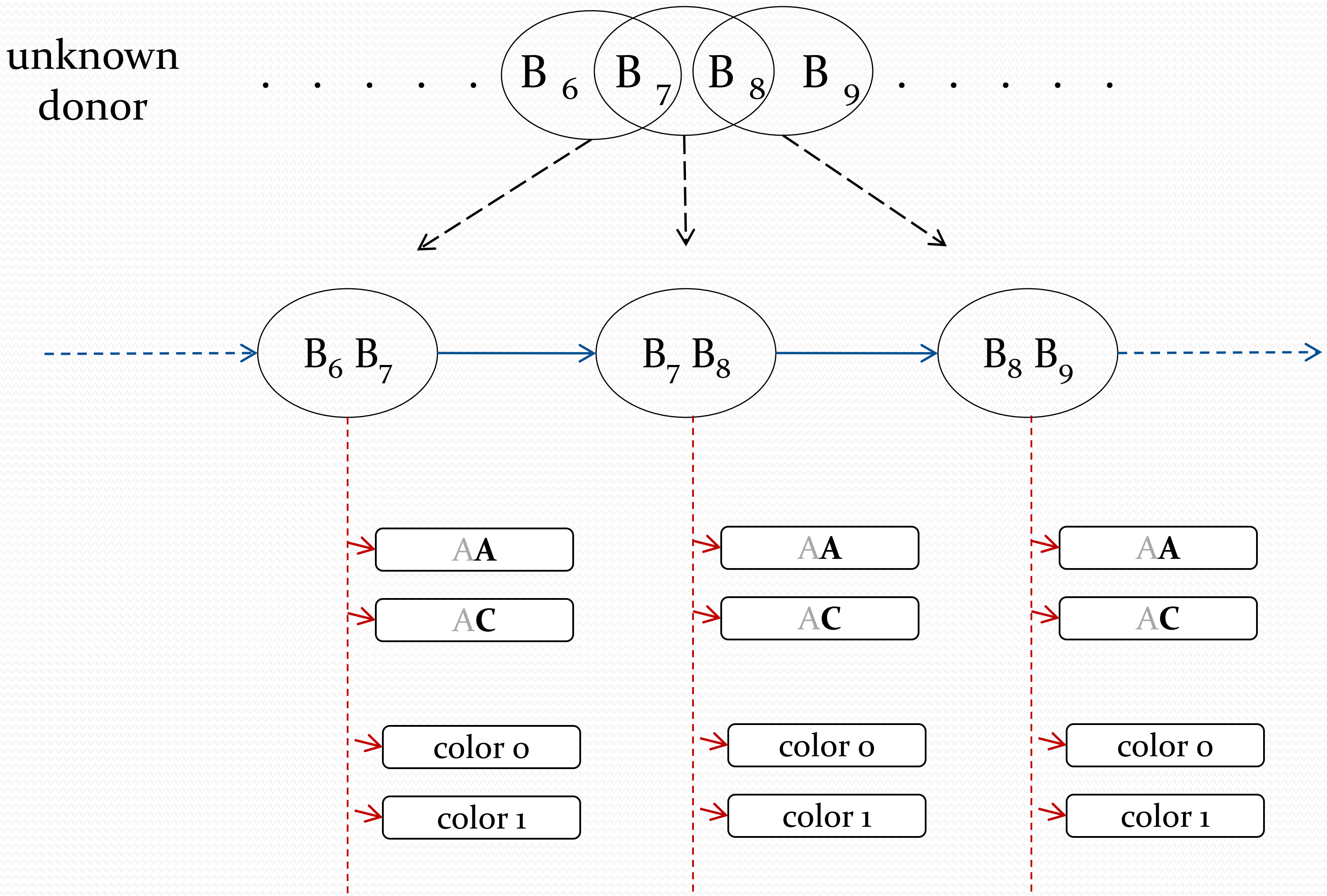


Hidden Markov Model (HMM)

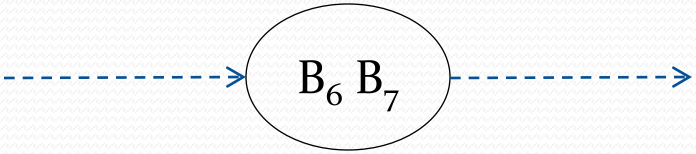
Apply HMM to variation detection:

- we don't know the state (donor), but
- we can observe some output determined by the state (aligned reads)

Hidden Markov Model (HMM)



States



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

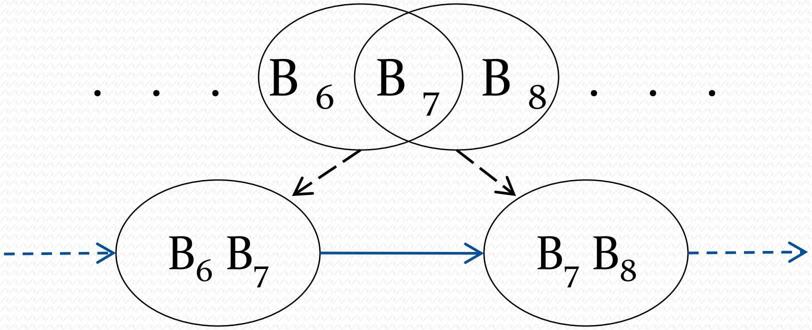
  :
- letters: TT color 0

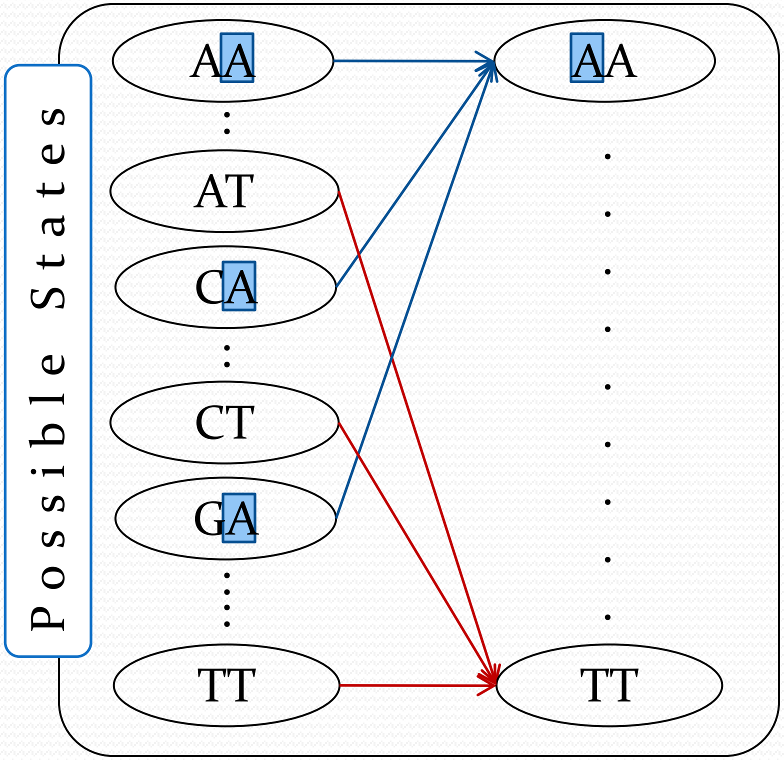
16 combinations

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

Why **pairs of letters**? Handle colors.
- AA and TT gives the same colors. Can't just model colors

Transitions





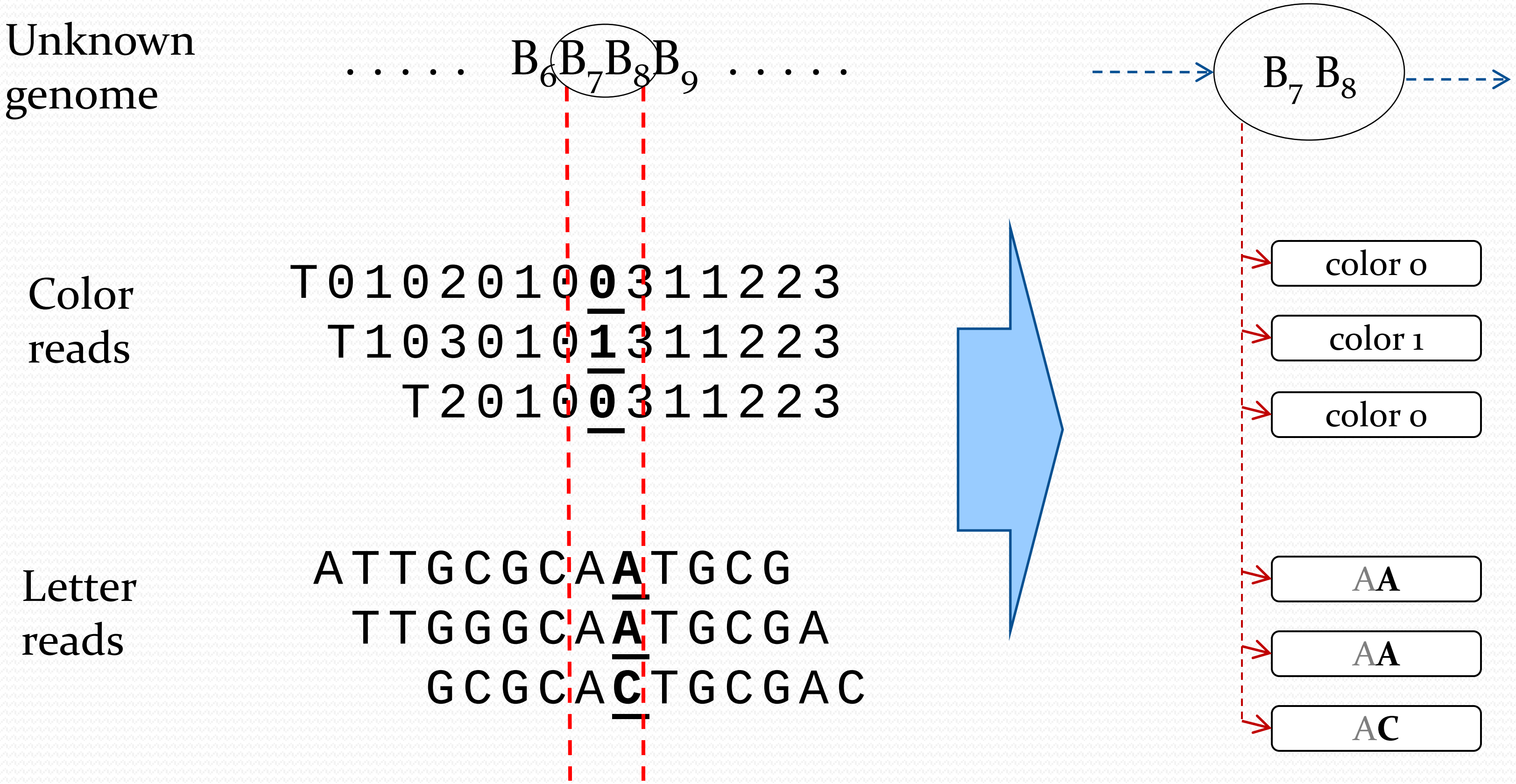
States

- 16 possible states
- only look at second letter

Transitions

- only certain transitions allowed
- when allowed, $p(X_t|X_{t-1}) = freq(X_t)$
- each state depends only on the previous states (Markov Process)

Emissions



Unknown genome

. . . . . $B_6 B_7 B_8 B_9$ . . . . .

Color reads

```
T01020100311223
 T10301013311223
    T2010031 1223
```

Letter reads

```
ATTGCGCAATGCG
 TTGGGCAATGCGA
   GCGCACTGCGAC
```

$B_7\ B_8$

color 0

color 1

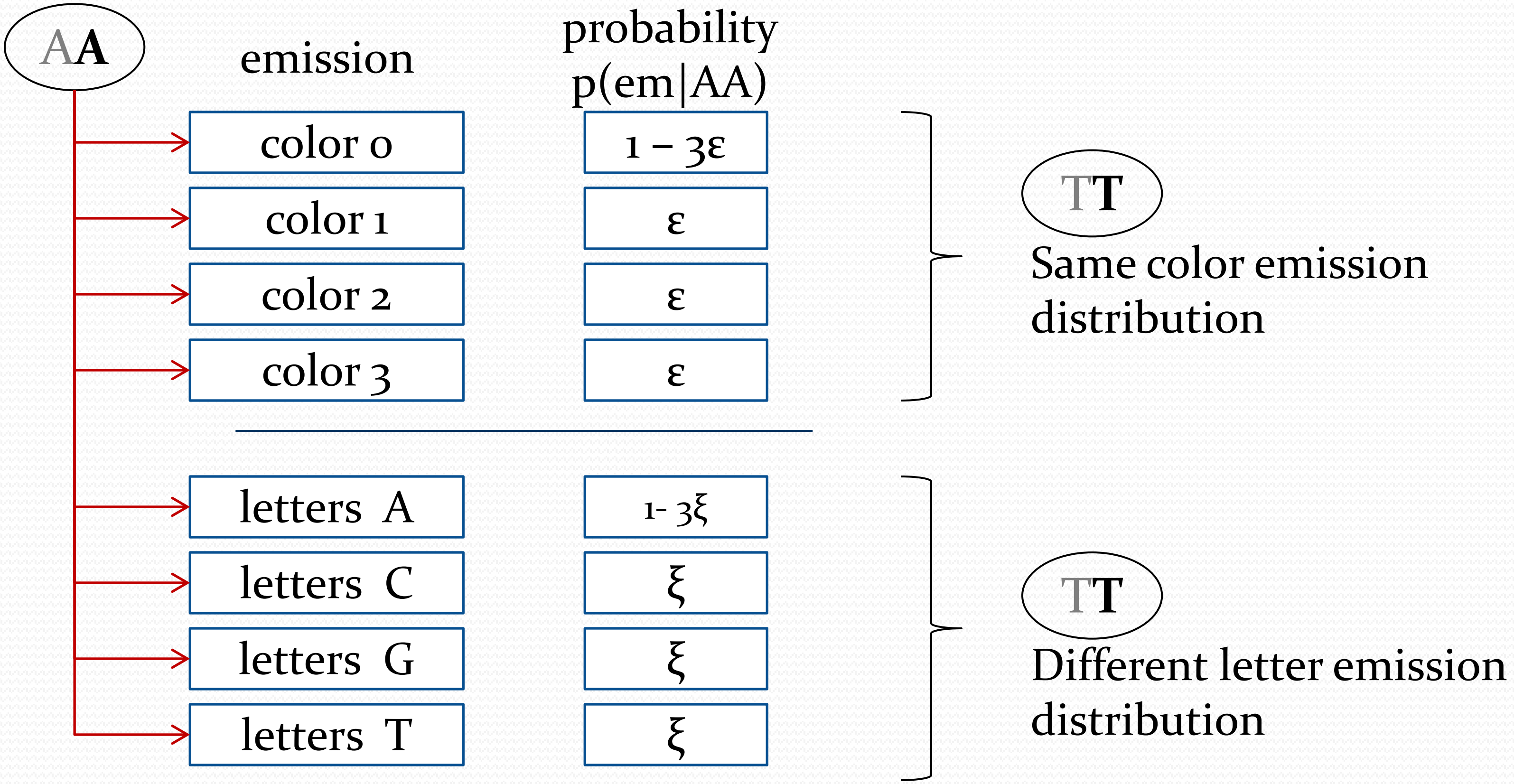
color 0

AA

AA

AC

Emission Probabilities



AA

emission	probability p(em|AA)
color 0	1 – 3ε
color 1	ε
color 2	ε
color 3	ε
letters A	1- 3ξ
letters C	ξ
letters G	ξ
letters T	ξ

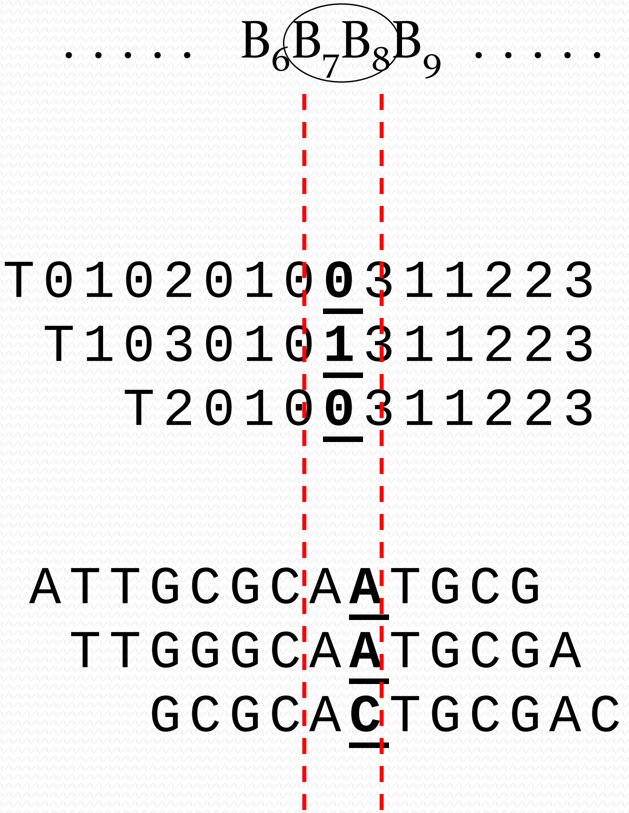
TT

Same color emission distribution

TT

Different letter emission distribution

Emission Probabilities



```
T01020100311223
 T103010131122 3
   T20100311223
```

```
ATTGCGCAATGCG
 TTGGGCAATGCGA
    GCGCACTGCGAC
```

Combining emission probabilities

- probability that this state emitted these reads.

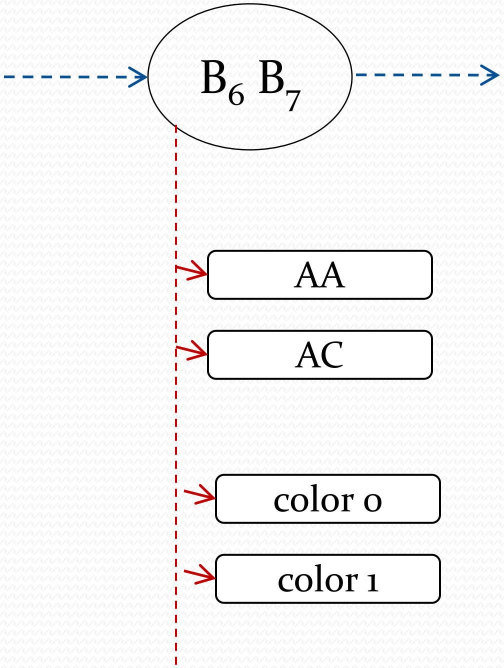
E.g. For state CC:

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

Simple HMM

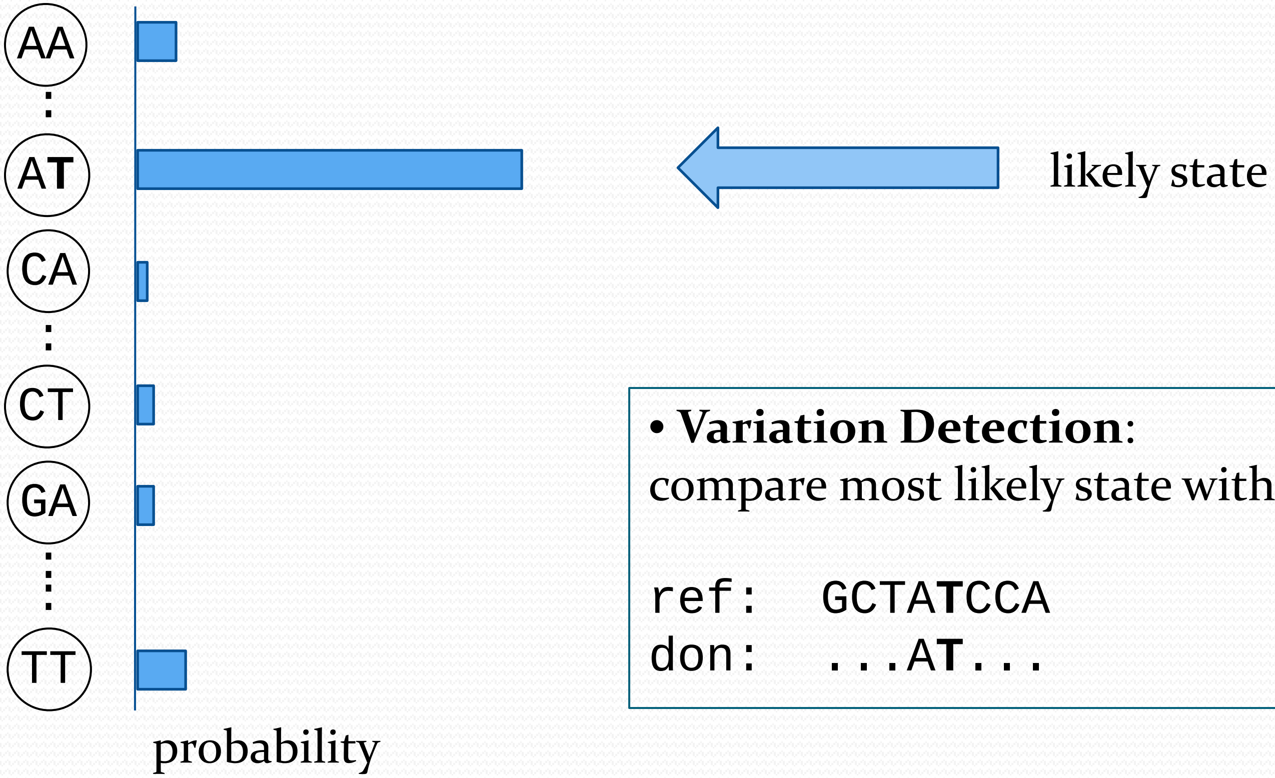
Summary

- unknown state
    - donor pair at location
- transitions
    - transition probabilities
- emissions
    - reads at location
    - emission probabilities



Forward-Backward Algorithm

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



- **Variation Detection**:

compare most likely state with reference:

```
ref:  GCTATCCA
don:  ...AT...
```

Motivation

Methods

Simple HMM Model

states, emissions, transitions, FB

Extended HMM Model

gaps, diploids, exceptions

Results

Summary

Extended HMM

Simple HMM

- only detects **homozygous SNPs**

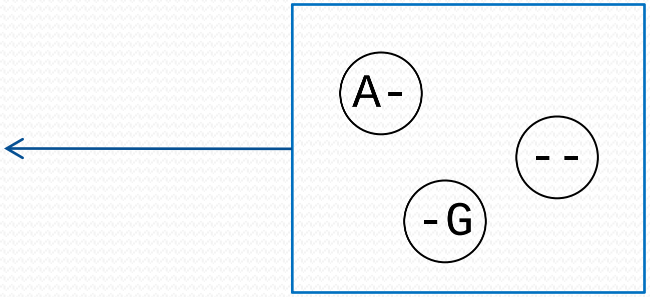
Extended HMM:

- short indels
- heterozygous SNPs
- complex error profiles & quality values

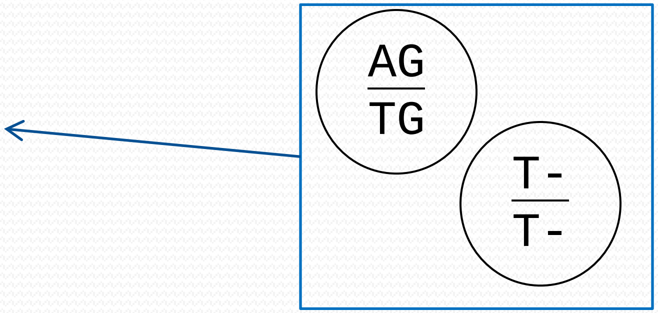
Expansion: Gaps and heterozygous SNPs

Expand states

- Have states that include gaps
    - emit: gap or color



- Have larger states, for diploids



- Transitions built in similar fashion as before
- Same algorithm, but in all we have 1600 states with very sparse transitions

Expansion

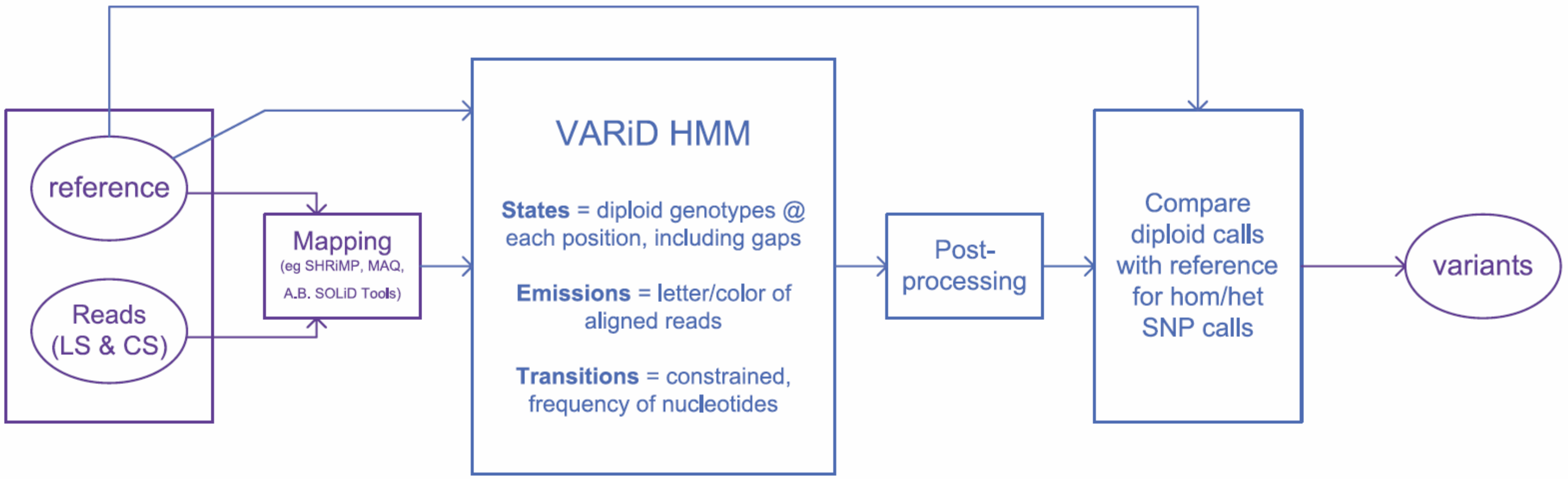
- Emission probabilities
  - Support quality values
  - Use variable error rates for emissions

- Translate through the first color
  - first color is incorrect
  - letter-space signal

```
Donor: ACAGCATCGGCATCGACTGC
        1123132303123321213
read: >T21231323031233212 13
       > C1231323031233212 13
```

- Post-process putative SNPs
  - correlated adjacent errors may support het SNPs
  - check putative SNPs

Summary


reference
Reads
(LS & CS)
Mapping
(eg SHRiMP, MAQ,
A.B. SOLiD Tools)
VARiD HMM
States = diploid genotypes @ each position, including gaps
Emissions = letter/color of aligned reads
Transitions = constrained, frequency of nucleotides
Post-processing
Compare diploid calls with reference for hom/het SNP calls
variants

blue: varid steps

Motivation

Methods

Results

Summary

Results

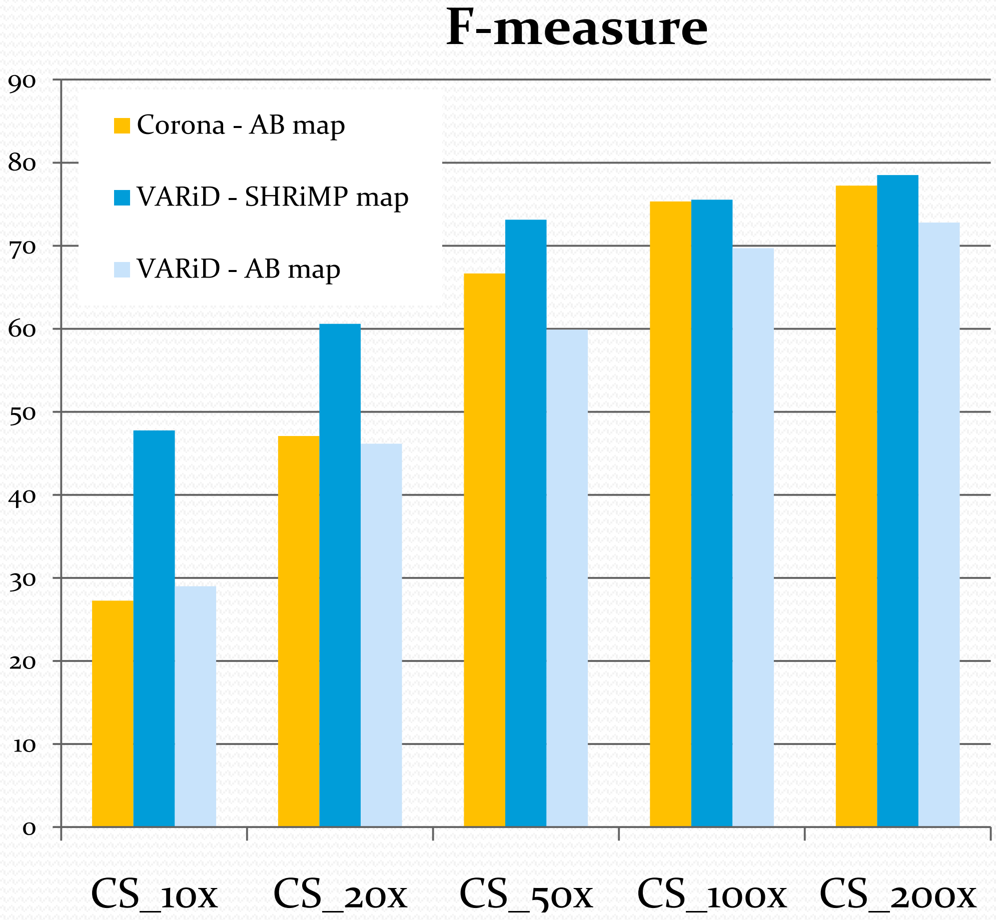
- Human dataset from Harismendy et al, 2009. (NA17156,17275,17460,17773) 454, SOLiD, Sanger

Color-space dataset:

- Compare random subsets:
  - Corona (with AB mapper)
  - VARiD (with SHRiMP)
  - VARiD (with AB mapper)

Conclusions:

- the three pipelines perform very similarly.
- High-coverage results is as good as can be achieved



Results

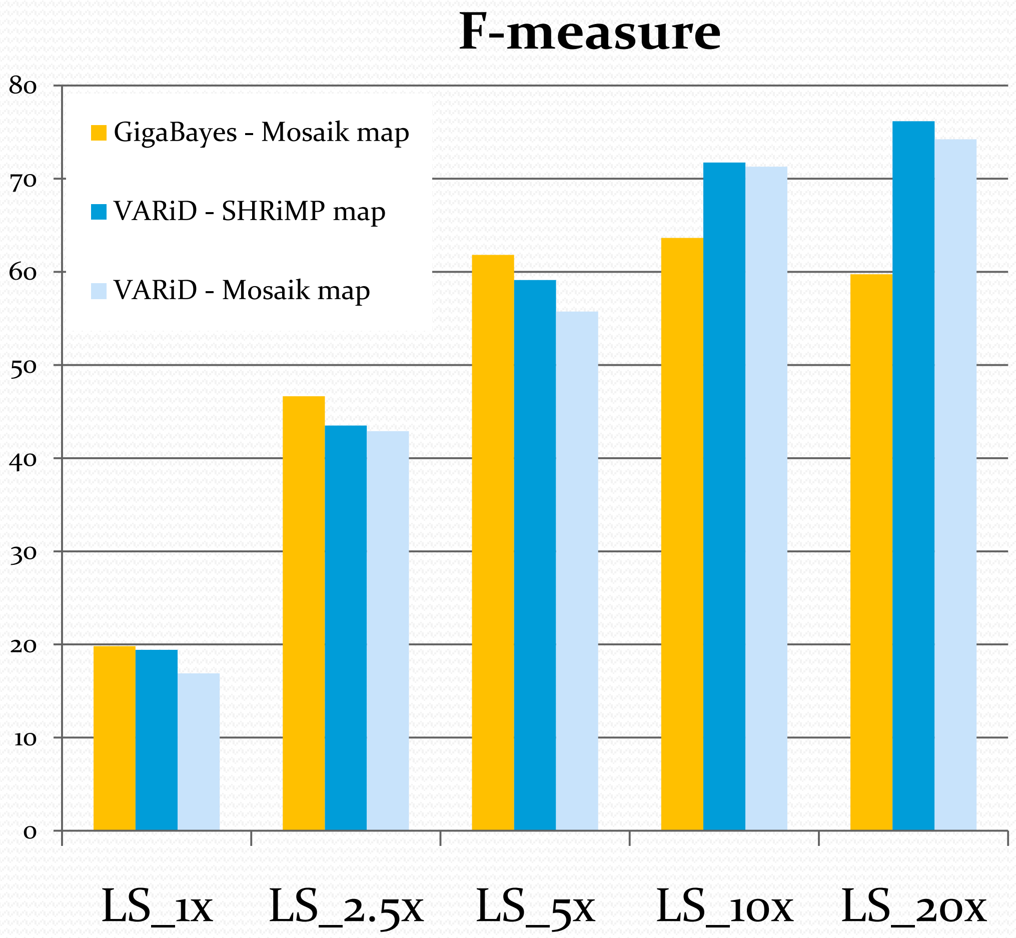
Letter-space dataset:

- Compare random subsets :
  - GigaBayes (with Mosaik)
  - VARiD (with SHRiMP)
  - VARiD (with Mosaik)

Conclusion:

•the three pipelines perform very similarly.

- High-coverage results is as good as can be achieved

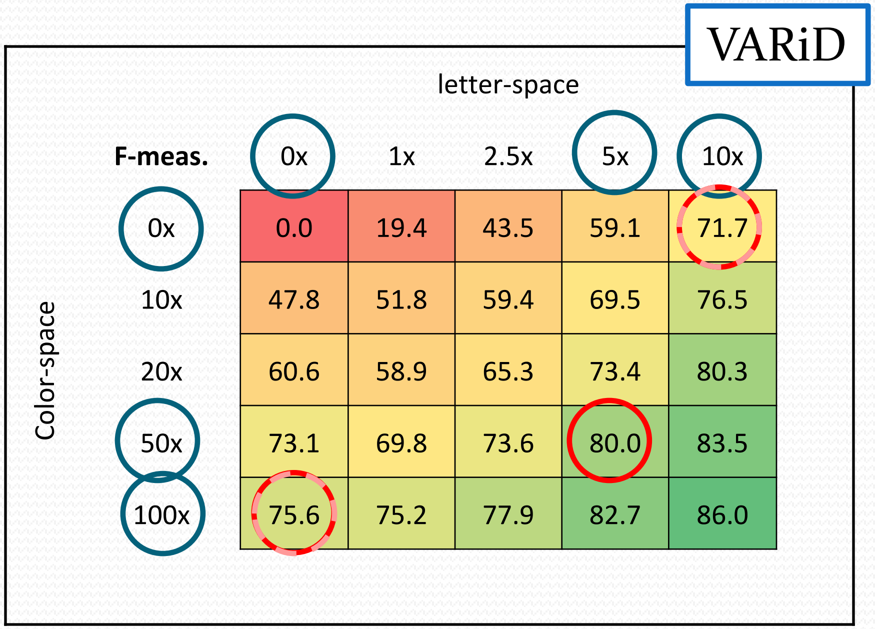


Results

VARiD Motivation:
Combining Letter-space and Color-space data to achieve increased accuracy in at-cost comparison

Assuming a same-cost comparison of:

- 10x letter-space (LS)
- 100x color-space (CS)
- 5x LS and 50x CS



VARiD

		letter-space				
Color-space	F-meas.	0x	1x	2.5x	5x	10x
	0x	0.0	19.4	43.5	59.1	71.7
	10x	47.8	51.8	59.4	69.5	76.5
	20x	60.6	58.9	65.3	73.4	80.3
	50x	73.1	69.8	73.6	80.0	83.5
	100x	75.6	75.2	77.9	82.7	86.0

Motivation

Methods

Results

Summary

Summary

Summary of VARiD

- HMM modeling underlying donor
- Treats color-space and letter-space together in the same framework
- no translation – take advantage of each technology's properties
- accurately calls SNPs, **short indels** in both color- and letter-space
    - improved results with hybrid data.

- **Website**: http://compbio.cs.utoronto.ca/varid

(VARiD freely available)

- **Contact**: varid@cs.utoronto.ca

Acknowledgements

Acknowledgements

- Steven Rumble, Sam Levy, Michael Brudno
- Misko Dzamba

Funding

MITACS

NSERC CRSNG

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NATIONAL INSTITUTES OF HEALTH

life technologies™

Partial Participant Support Award :

- ISCB, Department of Energy, and National Science Foundation

- **Website**: http://compbio.cs.utoronto.ca/varid
(VARiD freely available)
- **Contact**: varid@cs.utoronto.ca