VARiD: A variation detection framework for color-space and letter-space platforms

Adrian V. Dalca[1,2,*], Stephen M. Rumble[3], Samuel Levy[4] and Michael Brudno[2,5]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA, [2]Department of Computer Science, University of Toronto, Toronto, ON, Canada, [3]Department of Computer Science, Stanford University, Stanford, [4]Scripps Genomic Medicine, The Scripps Research Institute, La Jolla, CA, USA and [5]Donnelly Centre and the Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada

ABSTRACT

Motivation: High-throughput sequencing (HTS) technologies are transforming the study of genomic variation. The various HTS technologies have different sequencing biases and error rates, and while most HTS technologies sequence the residues of the genome directly, generating base calls for each position, the Applied Biosystem's SOLiD platform generates dibase-coded (color space) sequences. While combining data from the various platforms should increase the accuracy of variation detection, to date there are only a few tools that can identify variants from color space data, and none that can analyze color space and regular (letter space) data together.
Results: We present VARiD—a probabilistic method for variation detection from both letter- and color-space reads simultaneously. VARiD is based on a hidden Markov model and uses the forward-backward algorithm to accurately identify heterozygous, homozygous and tri-allelic SNPs, as well as micro-indels. Our analysis shows that VARiD performs better than the AB SOLiD toolset at detecting variants from color-space data alone, and improves the calls dramatically when letter- and color-space reads are combined.
**Availability:** The toolset is freely available at http://compbio.cs.utoronto.ca/varid
Contact: varid@cs.toronto.edu

1 INTRODUCTION

High-throughput sequencing (HTS) technologies are revolutionizing the way biologists acquire and analyze genomic data. HTS machines, such as 454/Roche, Illumina/Solexa and AB SOLiD are able to sequence up to a full human genome per week, at a cost hundreds fold less than previous methods. The resulting data consists of reads ranging in length between 35 and 400 nt, from unknown locations in the genome. Analysis of these datasets poses an unprecedented informatics challenge due to the sheer number of reads that a single run of an HTS machine can produce, the shortness of the reads, and the various technologies' different sequencing biases and error rates. The two basic steps in the discovery of variants in the human population from reads coming from any of these technologies are: first, the mapping of reads to a finished (reference) genome, and second the identification of variation by analysis of these mappings.

In the last few years, there have been many approaches proposed for mapping reads from HTS technologies (Campagna *et al.*, 2009; Langmead *et al.*, 2009; Li and Durbin, 2009; Li *et al.*, 2008a, b, 2009; Lin *et al.*, 2008; Rumble *et al.*, 2009 among many others; see

---

*To whom correspondence should be addressed.
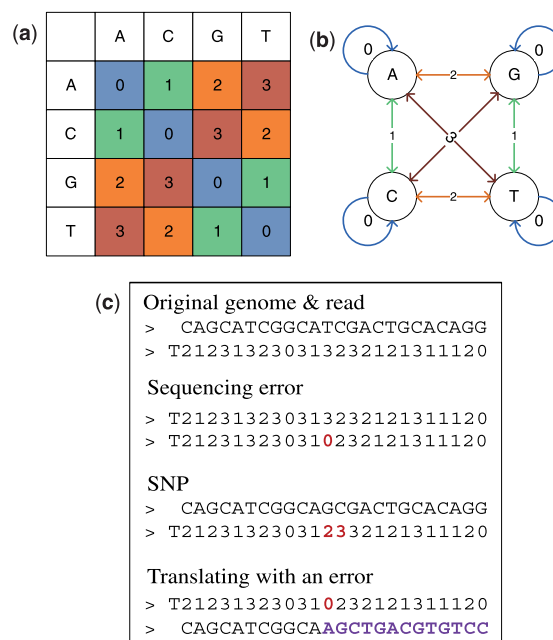
Fig. 1. Color-space description: Parts (**a**) and (**b**) show the correspondence between di-nucleotides and their color space representation with a translation matrix and the corresponding Finite State Automaton. In part (**c**), we show the effect of SNPs on the color-space representation of the read, as well as effect of sequencing errors on the trivial translation of the read from color to letter space. The first letter shown in the reads is actually the last letter of the linker, which helps us 'lock-in' on one of the four possible translations of a color-space read.

Dalca and Brudno, 2010; Flicek and Birney, 2009 for reviews) that utilize a wide variety of approaches. Compared to this multitude of mapping tools, there have only been a handful of toolsets for single nucleotide polymorphism (SNP) and small (1–5 bp) indel discovery. The main challenge in detecting these variants is using the error rates of the sequencing platform, the potentially incorrect mappings, and the varying coverage to determine the likelihood that a position represents a heterozygous or homozygous variant with respect to some reference genome. We use the term heterozygous to refer to the case when a single donor allele differs from the reference, and homozygous to refer to the case when both donor alleles differ from the reference, and are the same as each other. Tri-allelic SNPs, when the two donor alleles differ from each other and from the reference,

are rare. This variation detection task is further complicated by the different types of errors and data representation methods used by various technologies. For example, while the predominant error type in Illumina sequencing is the misreading of a base pair, in 454/Roche the most common mistake is insertion/deletion errors in a homopolymer (same base repeating multiple times). The AB SOLiD system introduced a dibase sequencing technique, where two nucleotides are read at every step of the sequencing process together as one *color*. Only four dyes are used for the 16 possible dibases (Fig. 1a), and the predominant error is the miscall of a color (colors are usually written as numbers 0–3). Most tools for variation detection (Li *et al.*, 2008a, 2009; Marth *et al.*, 1999) combine a detailed data preparation step, in which the reads are filtered, realigned and often rescored, with a nucleotide or heterozygosity calling step, typically done using a Bayesian framework. The typical parameters considered are the sequencing error rate, the SNP rate in the population (the prior) and the likelihood of misalignment (mapping quality). Most of the tools for SNP calling analyze one base of the reference genome at a time and do not use adjacent locations to help call SNPs (positions are considered independent).

AB SOLiD's dibase sequencing presents several unique challenges for SNP and indel identification. While typical, letter-space reads represent the DNA sequences directly as a string of A's, C's, G's and T's, one can think of dibase encoding as the output of a Finite State Automaton: consider each color as the shift from one letter to the next, so even though only four colors are generated, we can derive each subsequent letter if we know the previous one (Fig. 1b). Sequencing starts at the last letter of the molecule that connects to the DNA (the linker), which is known, thus enabling the translation of the whole read from color space into letter space. It is important to note, however, that if one of the colors in a read is misidentified (e.g. due to a sequencing error), this will change all of the subsequent letters in the translation (Fig. 1c). For this reason, simply translating the reads to letter-space would be impractical. While this error profile may at first seem detrimental, it can actually be advantageous when we need to decide if a particular difference between a read and the reference genome is due to an underlying change in DNA or a sequencing error: all SNPs will change two adjacent colors, while the probability that two adjacent colors are both misread is small, as error probabilities at adjacent positions are independent. Simultaneously, non-SNP genomic variants (e.g. polymorphisms at adjacent residues and micro-indels) have more complicated color-space signatures, complicating variation discovery.

Some tools for color-space SNP calling first map the reads in color space by translating the reference, but then translate the multiple alignment back to nucleotide space in order to call SNPs (Li and Durbin, 2009; Li *et al.*, 2008a). McKernan *et al.*, 2009 describe Corona Lite, a consensus technique where each valid pair of read colors votes for an overall base call. Currently, there are no methods that can simultaneously take full advantage of both color- and letter-space data to call variants—an important, consideration since the advantages and disadvantages of the various platforms are quite disparate. By combining these data sources, it is possible to exploit the strengths of multiple HTS technologies to improve on the accuracy of current SNP callers. Here, we present VARiD—a probabilistic approach for variant identification from either or both letter- and color-space data simultaneously. We represent both types of data as emissions from a hidden Markov model (HMM),

while the underlying genotypes of the sequenced genome are the hidden states. By applying the forward–backward algorithm on the HMM we generate, for every base of the genome, a probability distribution over the possible bases. In our testing, VARiD performs more accurately than AB's Corona Lite pipeline for just color-space data, while its ability to incorporate letter-space data allows for more accurate determination of genomic variants using multiple read types, simultaneously.

2 ALGORITHMS

In this section, we introduce our application of a HMM to the process of detecting variation from mapped reads. We begin by describing a simplified version of the model, and then describe the details of the full model and pipeline.

2.1 A hidden Markov model for variation detection

An HMM is a statistical model where the states of the system are hidden—that is, not observable directly—and respect a Markov progression. The observables are emissions from the hidden states. For a detailed introduction to HMMs, we refer the reader to Chapter 3 of Durbin *et al.* (1999). The structure of an HMM is defined in terms of the possible hidden states and the permitted transitions and hidden states and the permitted transitions between these. The model is parameterized by the emissions and transition probabilities. In the context of variation detection, we define the following HMM model (illustrated in Fig. 2):

- *States*: the unknown states in the HMM indicate the possible donor genotypes at each position in the genome. As we will model color-space, as well as letter-space data, and color-space sequencing corresponds to the change between adjacent nucleotides, the HMM will have states that correspond to *pairs* of consecutive positions. Overall, there are 16 possible states: {AA, AC, AG, AT, CA, …, TG, TT}, illustrated in green in Figure 2a.

- *Transitions*: as each state corresponds to a pair of nucleotides, two adjacent states will overlap by one nucleotide: for example, the state at positions (5, 6) will be followed by the state at positions (6, 7), thus sharing the nucleotide at position 6.
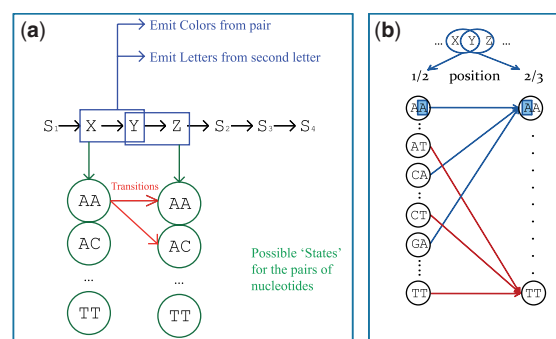
Fig. 2. Illustration of the simplified VARiD HMM. In (**a**) emissions, states and transitions are illustrated, and in (**b**) we illustrated in detail how one can transition from one state to the next. Note that Y is shared in the illustration (b), and hence we can only transition from a state ending in, say, letter A to a state starting with A.
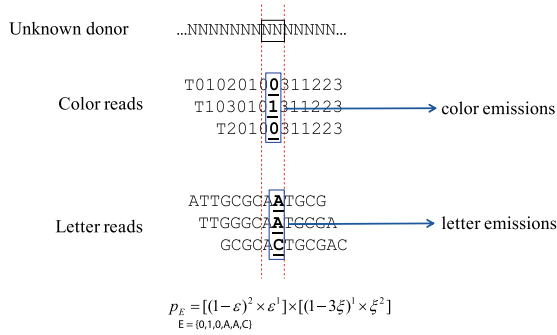
Fig. 3. This figure illustrates the concept of emissions in our problem: at the top, we have two adjacent positions in the unknown genome. We also have six aligned reads—three color-space, three letter-space. The exact aligned colors to this pair, and the exact aligned letters to the second letter in this pair represent the six emissions observed for this state. We can proceed to compute the probability that these emissions came from a state AA, AC, …. We show such a computation for the state CC. This example is also described in the text, see Equation (5).

Fig. 4. Possible emissions of the states AA and TT, with the respective probabilities. Here, ϵ and ξ are the error probabilities in color space and letter space. In the complete VARiD model, these errors will vary with their position in a read.
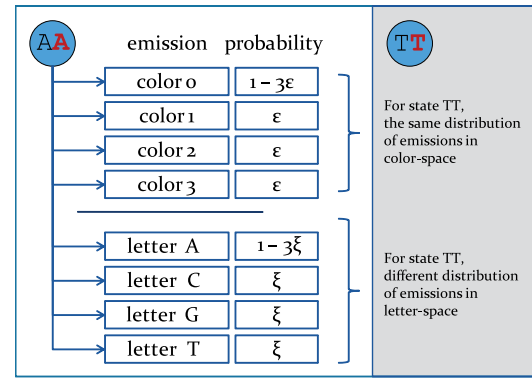
Consequently the transitions are constrained so that states that end with some nucleotide Y can only transition to states that start with the same nucleotide Y, thus forcing transitions that obey the overlap between adjacent states (Fig. 2b). Using this constraint and the frequency of each nucleotide, we define our *transition probabilities*:

$$P(\text{transition SZ} \rightarrow \text{XY}) = \quad (1)$$

$$p(XY|SZ) = \begin{cases} \text{frequency}(Y) & \text{if X} = \text{Z} \\ 0 & \text{otherwise} \end{cases}$$

For example, the state (TA) will have probability of 0 to transition in any state not starting with A due to our constraint, and the probability of transition to state (AY), where Y is one of {A, C, G, T}, is equal to nucleotide Y's frequency.

- *Emissions*: given that the states of the model correspond to the donor genotypes, the emissions are the donor reads at these loci, generated by either letter- or color-space sequencing technologies (Fig. 3). The genotype state at some position $(\rho, \rho+1)$ can emit one color and one letter (we arbitrarily choose the second, $\rho+1$ letter as the emission). As the states overlap, the first nucleotide is emitted by the previous state. Since the emissions are (mapped) reads, and since platforms and mappers are prone to error, a state corresponding to the di-nucleotide **CA** will emit color 1 with high probability, although it may emit other colors with some error probability ϵ. Similarly, **CA** will emit the letter A with high probability, but may emit other letters with some error ξ. We define the probability of emitting one particular color c or letter ℓ from the state **CA** as (Fig. 4):

$$P(\text{emission} = c | \text{state} = \text{CA}) = \quad (2)$$

$$q(c|\text{CA}) = \begin{cases} 1 - 3\epsilon & \text{if } c \text{ is } 1 \\ \epsilon & \text{if } c \text{ is } 0, 2 \text{ or } 3 \end{cases}$$

and

$$P(\text{emission} = \ell | \text{state} = \text{CA}) = \quad (3)$$

$$q(\ell|CA) = \begin{cases} 1 - 3\xi & \text{if } \ell \text{ is A} \\ \xi & \text{if } \ell \text{ is C, G or T} \end{cases}$$

Similar emission probabilities follow for all states. Since in general more than one read will cover a position, and we may have reads from different technologies, we combine the above definitions to get the emission probabilities for our HMM:

$$P(\text{emissions} = E | \text{state} = s) =$$

$$q(E|s) = \left(\prod_{\text{colors } c \in E} q(c|s) \right) \left(\prod_{\text{letters } \ell \in E} q(\ell|s) \right) \quad (4)$$

where E is a set of letter and color emissions at that position. For the example illustrated in Figure 3,

$$P(\text{emissions} = \{0,0,1,A,A,C\} | \text{state} = \text{CC}) =$$

$$\left((1-3\epsilon)^2 \epsilon^1 \right) \left((1-3\xi)^1 \xi^2 \right) \quad (5)$$

- *Genotyping*: we formulate the problem of variation detection from letter- and color-space sequencing as the problem of finding the maximum likelihood state for each genotype's position, given the emissions generated by the HMM. To obtain the most likely state at each position we use the forward–backward algorithm. This algorithm first computes, for each state, the probability of being in this state having observed all of the emissions prior to this position (forward probability), and the probability of starting in this state if we are to observe all the remaining emissions (backward probability). Combining the forward and backward probabilities for a specific location, one gets the overall likelihood of each state at that location given all the observed emissions (Fig. 5). We detect variants by comparing the most likely state with the reference nucleotide at this position.

2.2 VARiD: algorithm for variation identification

In the previous subsection, we described a simplified HMM for variation detection that can use both color- and letter-space data. This

simple HMM, however, calls only a single nucleotide per position, and cannot detect events such as micro-indels or heterozygous SNPs. In this section, we describe the full VARiD variation identification algorithm, including the expanded HMM utilized to address the above shortcomings, and the use of base and mapping quality values to parameterize the emission probabilities. We also describe the post-processing methods utilized in VARiD to filter some types of spurious calls. A summary of the VARiD pipeline and model is given in Figure 6.

2.2.1 Extensions to the HMM

- *Insertions and Deletions*: in order to detect micro-indels, the model must include gaps in the state definitions. Due to the nature of color-space sequencing, the expanded model needs to maintain the last letter before the current gap was started. For example, the `A--G` subsequence, represented by the states {(A –), (– –), (– G)}, should emit the color 2 of AG on the last state, which is accomplished by maintaining four gap types, gapA, gapC, gapG and gapT, with the rule that a gapX state can only follow the letter X or another gapX state. Thus, in addition to the 16 basic states there are also 24 gap states: 4 states (X, gapX), 4 states (gapX, gapX), and 16 states (gapX, Y), where X and Y are nucleotides {A, C, G or T}, giving a total of 40 states.
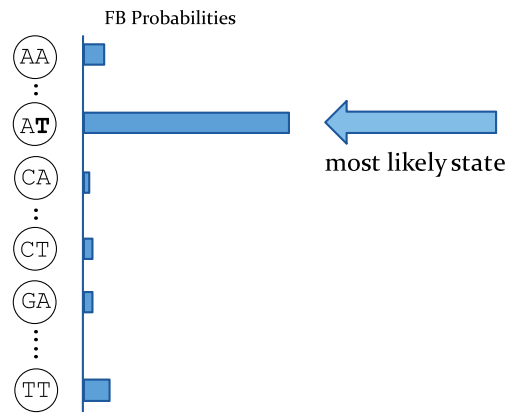

FB Probabilities

most likely state

Fig. 5. An example of the resulting probabilities given by the forward–backward algorithm: in this case, the state AT will be most likely and the nucleotide T will therefore be proposed.

These states allow for deletions with respect to the reference. The model requires no changes for insertions with respect to the reference (i.e. gaps in the reference), as the state sequence only describes the donor.

- *Heterozygous SNPs*: to allow for heterozygous variant detection, we build an expanded set of states by taking the cross-product of the state space with itself. Each state represents both alleles at a position and thus corresponds to a pair of dibases, e.g. (AC/AG) or (A-/TG). After expanding the states for indels and diploid states, there are a total of $40^2 = 1600$ states in the HMM. Similar to the transition probabilities above, only a small fraction of the possible transitions are allowed: states where the second nucleotides in the two alleles are A and G, for example, can only transition to states where the first nucleotides are A and G, and the transition probabilities in such cases are based on nucleotide frequencies. An example of resulting states and transitions is shown in Figure 7a.

- *Emission probabilities*: While the simple model described above used constant errors ϵ and ξ to parameterize color- and letter-space emissions, respectively, in practice the error rates vary with the position in the read, and most platforms also generate a quality score for each position in the read to indicate the likelihood of error. VARiD can use both of these sources of information, either converting a quality value into an error likelihood (assuming it is on the standard Phred scale) or using pre-specified error likelihoods for every position in a read. In the results presented below we use the second approach, as in our experience with the AB SOLiD data the quality values proved less informative than the read position. The per-position error frequencies are maximum likelihood estimates obtained from the alignments of the color-space reads. For the 454 data, we use a fixed error probability of 0.5%, also inferred from the mappings.

- *First color*: the first color in a color-space read is encoded relative to the last letter of the linker that connects the DNA to the slide. This will cause the first color in a read to be different from the corresponding color in other reads, which are encoded relative to the previous DNA letter. To address this, we 'translate through' the first color of the read, thus obtaining the first-sequenced DNA letter, and use this letter as an emission. For example, if a read began 'T2312…', it will be converted to
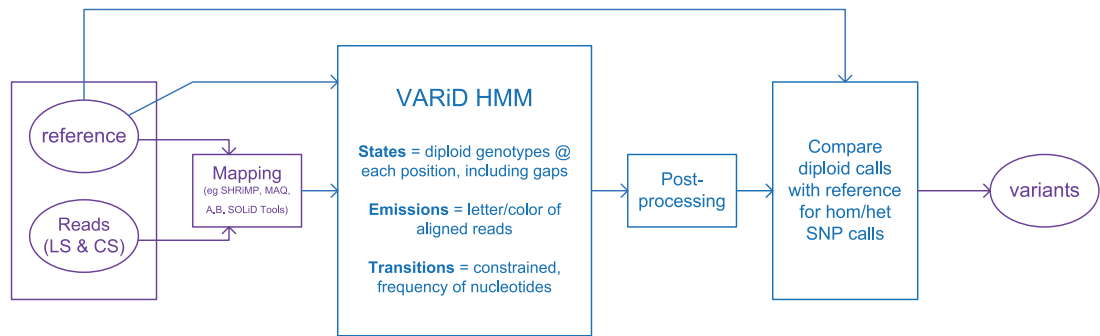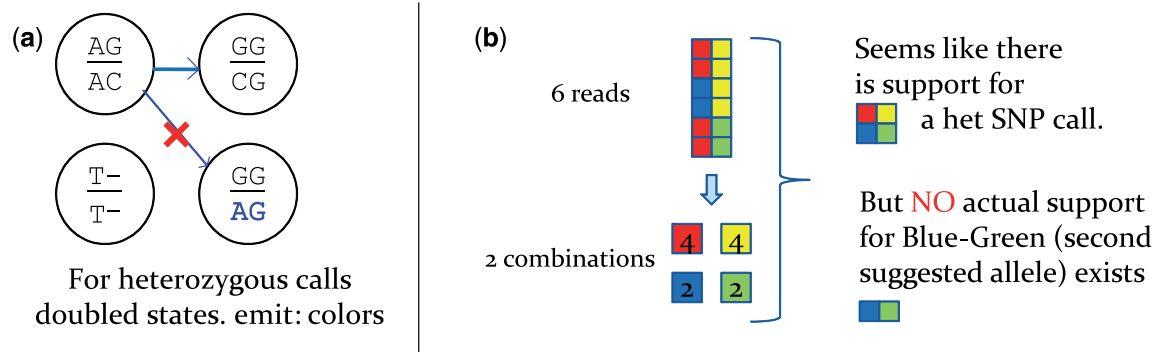

Fig. 6. A summary of the steps involved in the described pipeline. The purple sections are inputs, outputs or steps performed with previous software. The blue parts illustrate steps described in this manuscript.

Fig. 7. Diagram showing the expansions of the model. (**a**) we show examples of the expanded states that allow for gaps and heterozygous calls, as well as examples of allowed and not allowed transitions. (**b**) we note that adding a cleaning post-processing step is needed because of situations such as these: here we have six reads at two adjacent positions; when the colors of these reads are added up, it seems like we could call a heterozygous SNP represented by the allele combination such as red–yellow, blue–green, although the blue–green combination is actually not present in any read. Instead of incorporating a higher order model which would incur complexity costs, we simply check the (generally few) proposed SNPs and disregard cases such as these.

'C312…'. The 'C' character becomes the corresponding letter-space emission, while the remaining colors are unaffected. This modification allows VARiD to be used with color-space data only, by providing some letter-space emissions, as well as with letter- and color-space reads together.

2.2.2 Post-processing The HMM that VARiD utilizes is memoryless: the information about the specific reads that generated certain letters and colors is not maintained. This leads to the possibility that a valid path through the state space is not supported by any reads. Figure 7b depicts an example that may result in a heterozygous SNP prediction: four counts of red and two counts of blue for the first position, and four yellow, and two green for the second. Red:yellow and blue:green are considered 'valid' adjacent color changes that typically support a SNP. In this case, however, there are no individual reads that support the blue:green combination, indicating that this combination is actually unlikely to appear in the genome and hence is unlikely to be a heterozygous variant. While the proper approach to fixing this problem would be to use a higher order HMM, this would be computationally inefficient. We instead supplement the current probabilistic model with a post-processing step, where we verify that a statistically likely fraction of the reads directly support each heterozygous SNP call. This approach is fast, as putative SNPs are rare.

2.2.3 Running time The running time of the typical forward–backward algorithm is $O(nt)$, where n is the length of the sequence and t is the number of permitted transitions. While $t < k^2$, where k is the number of states, in the VARiD HMM $k = 1600$ and it is necessary to utilize sparse matrix operations to efficiently implement the forward–backward algorithm. Overall, the running time of VARiD is linear in the length of the genome. Furthermore, it is possible to parallelize VARiD over larger intervals by splitting the reference into smaller segments or windows, with the requirement that they be slightly overlapping. The overlapping regions can then be easily reconciled. VARiD required ∼4 min on a single Intel P4 Xeon 3.2GHz machine to predict variants in the 80 kb of the human genome that we analyze in the next section.

3 RESULTS

To test VARiD, we utilized the dataset of Harismendy *et al.* (2009), who sequenced several regions of the human genome, spanning a total of 260 kb, from four individuals (NA17156, NA17275, NA17460 and NA17773), both with the AB SOLiD platform and the 454/Roche Pyrosequencer. To validate the SNP calls, the authors also resequenced 80 kb from the same regions with Sanger sequencing. From the original high-coverage datasets, we generated reduced coverage, randomly selected subsets from the individuals with different degrees of coverage. To analyze the AB SOLiD data we ran the SOLiD System Analysis Pipeline Tool (Corona Lite 4.2.2 with the 35_3 schema) on the color-space data, as well as VARiD with both the AB Pipeline mappings as well as SHRiMP (Rumble *et al.*, 2009) mappings, for all of the read subsets. For the 454 data, we ran VARiD and gigaBayes (Marth *et al.*, 1999) on the letter-space reads (using Mosaik and SHRiMP as the mapping tools). Finally, we tested our prediction pipeline on various color- and letter-space subsets combined. We compared the variants called by each method with the Sanger validation set to compute the following statistics:

- Number of true positive (TP): SNPs that the predictor detects that are also in the validation set;
- Number of false positive (FP): SNPs the predictor calls variant that are not in the validation set;
- Precision: the number of true positives as a fraction of all predictions, $100*TP/(TP+FP)$;
- Recall: the fraction of true positives as a fraction of the validated SNPs, $100*TP/(TP+FN)$;
- F-measure: the harmonic mean of precision (P) and recall (R): $2*P*R/(P+R)$.

The results of our analysis are illustrated in Figures 8–10, where we present results of color space only, results of letter space only and results for combinations of the two sequence types, respectively.

In Figure 8, we present results from variation identification with VARiD and the Corona Lite SNP caller (http://www.solidsoftwaretools.com/gf/project/mapreads) using the color-space data. We ran VARiD both with the alignments produced

	VARiD - SHRiMP map					VARiD - AB map					Corona - AB map				
	Prec.	Rec.	F-meas	TP	FP	Prec.	Rec.	F-meas	TP	FP	Prec.	Rec.	F-meas	TP	FP
CS_10x	76.7	34.7	47.8	138	42	88.5	17.3	29.0	69	9	82.3	16.3	27.3	65	14
CS_20x	74.6	51.0	60.6	203	69	85.1	31.7	46.2	126	22	79.6	33.4	47.1	133	34
CS_50x	80.0	67.3	73.1	268	67	81.1	47.5	59.9	189	44	80.2	57.0	66.7	227	56
CS_100x	77.9	73.4	75.5	292	83	84.0	59.5	69.7	237	45	81.8	69.8	75.3	278	62
CS_200x	79.9	77.1	78.5	307	77	83.0	64.8	72.8	258	53	80.1	74.6	77.2	297	74

Fig. 8. Results illustrating performance of VARiD and Corona Lite on various coverage rates of color-space AB SOLiD reads. In the first of the three sections, we ran VARiD on various datasets aligned with the SHRiMP tool, in the second we ran it with AB mapper output and finally in the third we ran the Corona Lite pipeline on the AB mappings. In general, the results show that variation detection is difficult even with high coverage of color space, and the results are dependent on the coverage and the mapping package used—for example, VARiD with SHRiMP mappings tends to have slightly lower precision, but higher recall, leading to higher F-measure scores, especially at lower coverages, while VARiD with AB mappings has higher precision, but also lower recall.

	VARiD - SHRiMP map					VARiD - Mosaik map					GigaBayes - Mosaik map				
	Prec.	Rec.	F-meas	TP	FP	Prec.	Rec.	F-meas	TP	FP	Prec.	Rec.	F-meas	TP	FP
LS_1x	62.1	11.5	19.4	41	25	39.2	10.8	16.9	38	59	80.4	11.3	19.8	45	11
LS_2.5x	73.1	31.0	43.5	122	45	67.4	31.5	42.9	124	60	70.2	34.9	46.6	139	59
LS_5x	74.0	49.2	59.1	196	69	63.0	50.0	55.7	199	117	64.0	59.8	61.8	238	134
LS_10x	76.4	67.6	71.7	269	83	74.5	68.3	71.3	272	93	59.2	68.8	63.6	274	189
LS_20x	70.3	83.2	76.2	331	140	67.7	82.2	74.2	327	156	55.8	64.3	59.7	256	203

Fig. 9. Results of running VARiD (SHRiMP alignments), VARiD (Mosaik alignments) and gigaBayes (Mosaik alignments) on all individuals of our datasets, using the 454/Roche data at various coverages. VARiD with SHRiMP mappings and gigaBayes have similar precision and recall at lower (10×) coverage, while VARiD with Mosaik alignments performs slightly worse. However, at high coverage (20×), VARiD with SHRiMP mappings has 70% precision to gigaBayes' 56%, and has 83% recall to gigaBayes' 64%, thus showing overall improvement.

		letter-space				
Prec.		0x	1x	2.5x	5x	10x
Color-space	0x	0.0	62.1	73.1	74.0	76.4
	10x	76.7	78.2	78.8	80.5	80.2
	20x	74.6	76.9	81.8	82.4	81.2
	50x	80.0	81.1	82.8	83.8	83.8
	100x	77.9	80.1	82.8	84.7	84.0

		letter-space				
Recall		0x	1x	2.5x	5x	10x
Color-space	0x	0.0	11.5	31.0	49.2	67.6
	10x	34.7	38.7	47.7	61.1	73.1
	20x	51.0	47.7	54.3	66.1	79.4
	50x	67.3	61.3	66.3	76.6	83.2
	100x	73.4	70.9	73.6	80.7	88.2

		letter-space				
F-meas.		0x	1x	2.5x	5x	10x
Color-space	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

Fig. 10. These numbers show the improvements we can obtain when combining reads from various platforms. Comparing at cost, for example, we can look at combining 50× of AB SOLiD color-space data with 5× of 454/Roche data. Comparing to the equivalent cost of 454/Roche (10×) we achieve 7% more precision and 9% higher recall in the combined run. Similarly, comparing to the equivalent cost of AB SOLiD color-space data (100×), we obtain 6% better precision and 3% better recall. Another example can be found by looking at the CS-100× and LS-10× combination, and comparing with 200× of CS or 20× of LS in Figures 8 and 9.

by the AB pipeline for the Corona caller and with alignments generated by SHRiMP. While the results as a whole demonstrate the difficulty of calling variants from color-space data, even at high coverages, a direct comparison of the two SNP calling pipelines shows that at low-coverage (10×) VARiD outperforms the Corona pipeline when using the same set of mappings generated by AB's own mapping tool, while at higher coverage VARiD has better precision and worse recall (and a lower F-measure). The VARiD + SHRiMP pipeline has slightly lower precision than Corona and VARiD + AB mapper, but a significantly better recall, leading to a higher F-measure score.

In Figure 9, we compare results of running the VARiD framework on the 454/Roche letter-space data using the Mosaik alignments as well as using the SHRiMP alignments, compared to gigaBayes using Mosaik alignments. At low coverage (1–5×), the gigaBayes SNP caller produces the best results, having higher precision with similar recall. At higher coverages (10–20×), VARiD outperforms gigaBayes with higher recall and higher precision, regardless of the mapper used to generate the alignments.

Figure 10 shows the main advantage of the VARiD pipeline: its ability to combine color- and letter-space reads. In determining useful combinations of the SOLiD and 454/Roche subsets for running on the VARiD framework together, we considered the cost and accuracy of each platform. The 454/Roche contains a relatively high indel count, but has much more accurate base calls. At the same time, the 454 platform is ∼10 times more expensive. Therefore, we considered combining read coverages with 10-fold more AB SOLiD than 454 data. For example, we may combine 50× of color-space reads with 5× letter-space, giving us the equivalent of 100× of AB SOLiD or 10× of 454 in terms of cost. Of course, the best trade-off

will vary depending on the costs of the platforms and their respective accuracies.

In Figure 10, we consider the various possible coverage combinations between the AB SOLiD data and the 454/Roche. In general, the performance of VARiD on a certain coverage of color-space data can be greatly improved with just a small number of 454 reads. More concretely, comparing at cost we can look at $50\times$ coverage of color space with $5\times$ coverage of 454 data: when combined, we find 84% precision and 77% recall. Looking at the cost equivalent coverage of just 454 data—$10\times$—gives 7–9% lower precision and recall. Similarly, for the cost equivalent coverage of AB SOLiD data—$100\times$—will again perform worse. Combining the data thus shows significant improvement over predicting variation from letter or color space only.

4 DISCUSSION

The various HTS technologies that have emerged in the past few years have different data representations, advantages, biases and features. In this work, we introduced a novel probabilistic framework for variation identification, which can use both letter- and color-space data simultaneously. We have shown in our results that when using only color-space data—a data type for which very few genomic analysis tools exist—the model outperforms the AB SOLiD toolkit Corona Lite, and performs on par with gigaBayes predictions for letter-space data alone. More importantly, when the color- and letter-space data are combined, the VARiD framework allows for a significant performance increase, demonstrating that a method that can take into consideration multiple technologies, combining their different advantages and compensating for their different weaknesses can achieve higher accuracy variant predictions than are possible from any single data type.

ACKNOWLEDGEMENTS

We thank Adrian Dalca Sr for help with the implementation.

REFERENCES

Campagna,D. *et al.* (2009) Pass: a program to align short sequences. *Bioinformatics*, **25**, 967–968.

Dalca,A.V. and Brudno,M. (2010) Genome variation discovery with high-throughput sequencing data. *Brief Bioinform*., **11**, 3–14.

Durbin,R. *et al.* (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Flicek,P. and Birney,E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Meth.*, **6**(Suppl.11), S6–S12.

Harismendy,O. *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32+.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25+.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome res.*, **18**, 1851–1858.

Li,R. *et al.* (2008b) Soap: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Li,R. *et al.* (2009) Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lin,H. *et al.* (2008) Zoom! Zillions of oligos mapped. *Bioinformatics*, **24**, 2431–2437.

Marth,G.T. *et al.* (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.

McKernan,K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.*, **19**, 1527–1541.

Rumble,S.M. *et al.* (2009) Shrimp: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386+.