BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–6

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

Adrian V. Dalca [1,2,*], Stephen M. Rumble [2,3], 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, CA, USA
[4]Scripps Genomic Medicine, The Scripps Research Institute, La Jolla, CA, USA
[5]Banting and Best Department of Medical Research, University of Toronto, Toronto, ON, Canada

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Next Generation Sequencing (NGS) technologies are transforming the study of genomic variation. While most NGS 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. Di-base encoding is optimized for variation, and the various NGS technologies have different sequencing biases and error rates. Combining data from the various platforms should therefore increase accuracy of variation detection. Yet, 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-space and color-space reads simultaneously. VARiD is based on a Hidden Markov Model (HMM), and it allows for accurate detection of heterozygous, homozygous and tri-allelic SNPs, as well as microindels by running the Forward-Backward algorithm on the resulting HMM. Our analysis shows that VARiD performs better than the ABI SOLiD toolset at detecting variants from color-space, and improves the calls dramatically when letterspace and color-space reads are combined.

Availability: The toolset is freely available at http://compbio.cs.utoronto.ca/varid.

Contact: adalca@mit.edu; brudno@cs.toronto.edu

1 INTRODUCTION

Next Generation Sequencing (NGS) technologies are revolutionizing the way biologists acquire and analyze genomic data. NGS machines, such as 454/Roche, Illumina/Solexa, and ABI 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 35-400 nucleotides, from unknown locations in the genome. Analysis of these datasets poses an unprecedented informatics challenge, because of the sheer number of reads that a single run of an NGS machine can produce, because the reads are significantly shorter, and because the different technologies have very 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 NGS technologies (Lin *et al.* (2008), Li *et al.* (2008a), Li *et al.* (2008b), Li *et al.* (2009), Campagna *et al.* (2009), Langmead *et al.* (2009), Li and Durbin (2009), Rumble *et al.* (2009)), 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 for this task lies in judging the likelihood that a position is a heterozygous or homozygous variant given the error rates of the various platforms, the probability of bad mappings, and the amount of support or coverage. This is further complicated by the different types of errors and data representation methods used by the various technologies. For example, while the predominant error type in Illumina is the mis-reading of a basepair, in 454/Roche the most common mistake is insertion/deletion errors in a homopolymer (same base repeating multiple times). The ABI 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, and the predominant error is the miscall of a color (colors are often written as numbers 0-3). Most tools for variation detection (Marth *et al.* (1999), Li *et al.* (2008a), Li *et al.* (2009)) combine a detailed data preparation step, in which the reads are filtered, re-aligned, and often re-scored, 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, as they are usually independent in letter-space sequencing.

*to whom correspondence should be addressed

ABI SOLiD sequencing is different in this respect. 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 (see Figure 1). 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 (Figure 1). 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: most underlying variants in the DNA will change two adjacent colors (with some exceptions), while the probability that two adjacent colors are both misread is small, as error probabilities at adjacent positions are thought to be independent, and hence very small together.

AB SOLiD's di-base sequencing presents several unique challenges for SNP and indel identification. 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 *et al.* (2008a), Li and Durbin (2009)). 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 call SNPs from both color-space and letter-space data - 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 NGS technologies to improve on the accuracy of current SNP callers. In this paper we present VARiD - a probabilistic approach for variant identification from either or both letter-space 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 ABs' Corona Light 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.

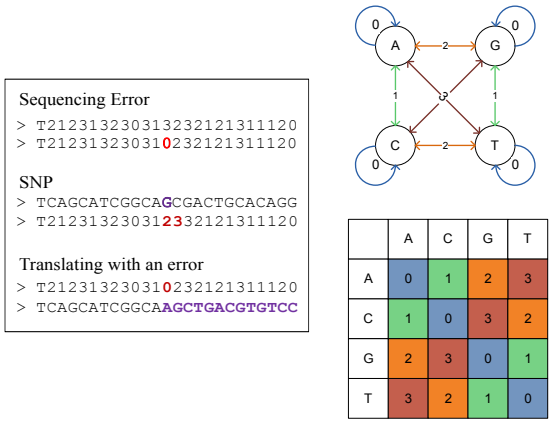

Fig. 1. Color-space description: On the left, we note the difference between a SNP and a color-space error, and the effects of translating a read with an error. The first letter shown in the reads is actually the last letter of the linker, which helps us "lock-in" on one of the four possible translations of a color-space read. On the right, we have the translation matrix and its associated Finite State Automaton.

2 ALGORITHMS

In this section we introduce our application of a Hidden Markov Model (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

A Hidden Markov Model 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 of HMMs with an application to computational biology, 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 between these. The model is then parametrized by the emission and transition probabilities. In the context of variation detection, we define the following HMM model (illustrated in Figure 2):

- **States.** The unknown states in the HMM indicate the possible donor genotypes at each position in the genome. Because 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.** Because 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. 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 (see Figure 2b). Using this constraint and the frequency of each nucleotide, we define our **transition probabilities**:

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

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

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 those loci, generated by either letter-space or color-space sequencing technologies (Figure 3). The genotype state at some position $(\rho, \rho+1)$ emits one color and one letter (we arbitrarily choose the second, $\rho+1$ emission). Because 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 CA will emit color 1 with high probability, although it may emit other colors, i.e. the reads may see other colors at this position, 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 or letter given a state by (also see Figure 3):

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

and

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

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:

$$P(\text{emissions} = E|\text{state} = s) =$$
$$q(E|s) = \left(\prod_{\text{colors } c \in E} q(c|s)\right)\left(\prod_{\text{letters } \ell \in E} q(\ell|s)\right). \tag{4}$$

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

$$P(\text{emissions} = \{0,0,1,\text{A},\text{A},\text{C}\}|\text{state} = \text{CC}) = \tag{5}$$
$$\left((1-3\epsilon)^2\epsilon^1\right)\left((1-3\xi)^1\xi^2\right)$$

- **Genotyping.** We formulate the problem of variation detection from letter-space 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. We can solve for these using the Forward-Backward algorithm, which yields the likelihood of each state at each location (Figure 5). We detect variants by comparing the most likely state with the reference nucleotide at this position.

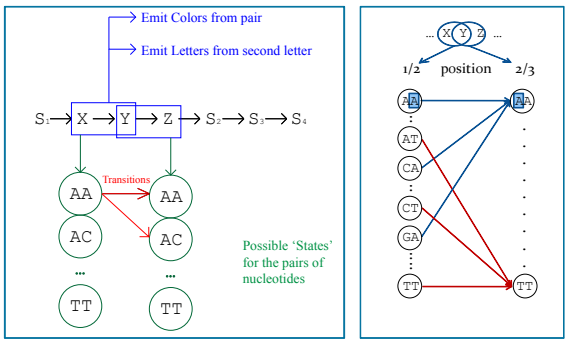

Fig. 2. Illustration of the VARiD HMM Model. On the left, emissions, states and transitions are illustrated, and on the right we illustrated in detail how one can transition from one state to the next. Note that Y is shared in the illustration on the right, and hence we can only transition from a state ending in, say, letter A to a state starting with A

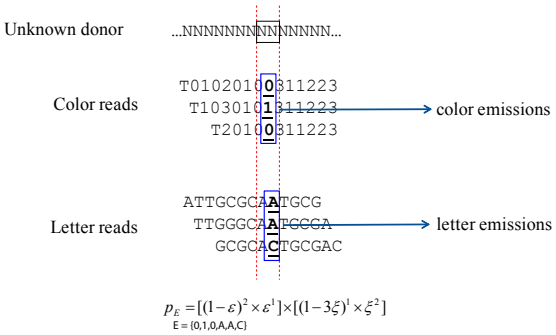

Fig. 3. This figure illustrates the concept of emissions in our problem: at the top, we have two adjacent positions in the unknown genome. We also have 6 aligned reads - 3 color-space, 3 letter-space. The exact aligned colors to this pair, and the exact aligned letters to the second letter in this pair represent the 6 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 eq (5).

2.2 VARiD: Algorithm for Variation Identification

In the previous sub-section we described a simplified HMM model for variation detection. 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 parametrize the emission probabilities. We also summarize the post-processing methods utilized in VARiD to filter some types of spurious calls.

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. Because of the nature of color-space sequencing, the expanded model needs to maintain the last letter before the current gap was started. For

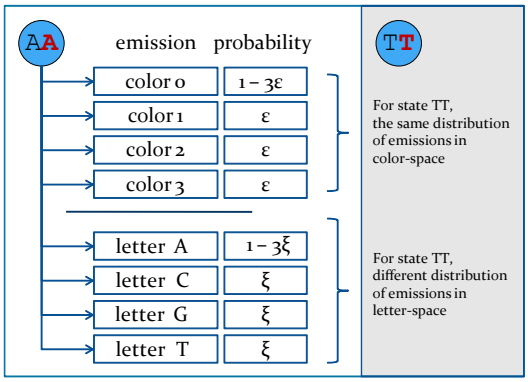

Fig. 4. Possible emissions of the states AA and TT, with the respective probabilities. Here, ϵ and ξ are the error probabilities in color-space and letter-space. In the complete VARiD model, these errors will vary with their position in a read.

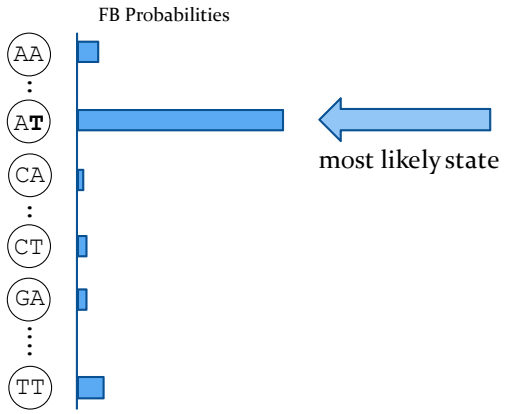

Fig. 5. An example of the resulting probabilities given by the Forward-Backward algorithm: in this case, the state AT will be most likely, and the nucleotide T will therefore be proposed.

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, as demonstrated in Figure 6a. 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.

- **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 (as demonstrated in the Figure 6b).

- **Emission probabilities.** While the simple model described above used constant errors ϵ and ξ to parametrize 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 PHRAP 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 ABI 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 ”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-space and color-space reads together.

A summary of the VARiD pipeline and model is given in Figure 7.

2.2.2 Post Processing The HMM that VARiD utilizes is memory-less: 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. For example, Figure 6b depicts an example that may be predicted as a heterozygous SNP: 4 counts of red and 2 counts of blue for the first position, and 4 yellow, and 2 green for the second are valid adjacent color changes. At the same time 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 position. While such cases are rare, we supplement the probabilistic model with a post-processing step where we verify that a statistically likely fraction of the reads directly support each heterozygous SNP call.

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 $\sim$ 4 minutes on a single Intel P4 Xeon 3.2GHz machine to analyze the 80kb regions 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 260kb, from four individuals (NA17156, NA17275, NA17460 and NA17773), both with the ABI SOLiD platform and the 454/Roche Pyrosequencer. We also obtained and utilized quality values for the color-space reads. To validate the SNP calls the authors also resequenced the same regions with Sanger sequencing. From the original high-coverage data sets, we generated reduced-coverage, randomly selected subsets from the individuals with different degrees of coverage. To analyze the ABI 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-space 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, $100 * TP/(TP + FP)$;
- the Recall: the fraction of true positives as a fraction of the validated set, $100 * TP/ValidatedSNPs$.

The results of our analysis are illustrated in Tables 1-3, where we present color-space only results, letter-space only results, and results for combinations of the two sequence types.

In Table 1 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 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 VARiD outperforms the Corona pipeline when using the same set of mappings generated by AB's own mapping tool. The VARiD + SHRiMP pipeline performs similarly to the VARiD + AB Mapper pipeling, having a higher recall rate, but simultaneously a lower precision.

In Table 2 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 coverages, the gigaBayes SNP caller produces the best results, having higher precision with similar recall. At higher coverages, however, VARiD outperforms gigaBayes with higher recall and higher precision.

Table 3 shows the main advantage of the VARiD pipeline: its ability to combine color-space 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, it can be estimated that the 454 platform is 10 times more costly. Therefore, we considered combining read coverages with 10-fold more ABI SOLiD than 454 data. For example we may

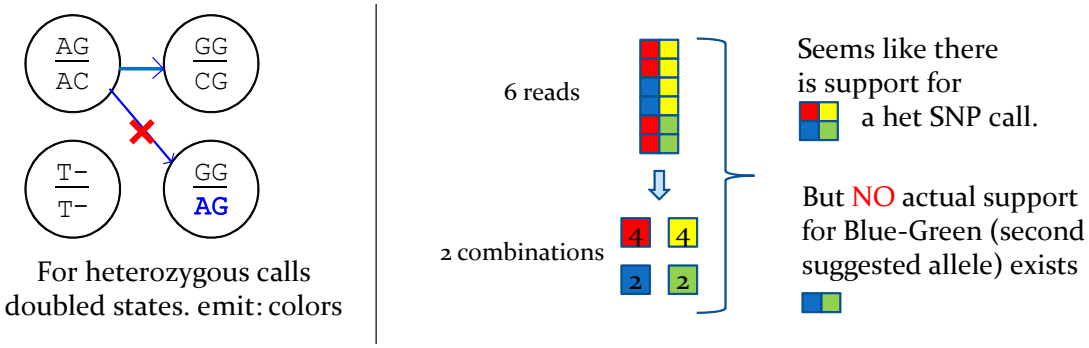

Fig. 6. Diagram showing the expansions of the model: on the left, expanding the states to allow for gaps and heterozygous calls, as well as examples of allowed and not allowed transitions. On the right, adding a cleaning post-processing step is needed because of examples such as these: here we have 6 reads at 2 adjacent positions: when the colors of these reads are added up, it seems like we could call a geterozygous SNP represented by the allele combinations (red-yellow), (blue-green), although the blue-green combination is actually not present in any read.

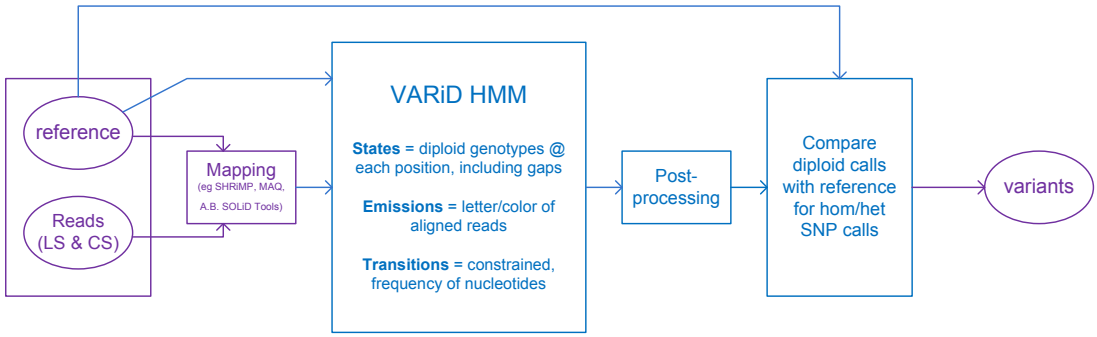

Fig. 7. A summary of the steps involved in the described pipeline. The purple sections are inputs, outputs or steps performed with previous software. he blue parts illustrate steps described in this manuscript.

combine 50x of color-space reads with 5x letter-space, giving us the equivalent of 100x of AB SOLiD or 10x of 454 in terms of cost. Of course the best tradeoff will vary depending on the costs of the platforms and their respective accuracies.

In table 3 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 50x coverage of color-space with 5x coverage of 454 data: when combined, we find 84% precision and 77% recall. Looking at the cost equivalent coverage of 454 data - 10x - VARiD gives around 7-9% lower percentage in both precision and recall, while GigaBayes precision will be even lower. Similarly, for the cost equivalent coverage in AB SOLiD data, 100x, VARiD and Corona will again perform worse as can be seen in the 100x row's first entry. Combining the data, therefore, shows significant improvement over predicting variation from letter-space or color-space only – 50x of color-space with 5x of 454 can perform better than 100x of color-space.

4 DISCUSSION

The various NGS technologies that have emerged in the past few years have different data representations, advantages, biases and features. In this work we introduced a novel probabilitic framework for variation identification which can use both letter-space 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 can perform on par or even better than the ABI SOLiD toolkit Corona Lite, and similarly can match or improve on gigaBayes predictions for letter-space data. More importantly, when the color-space 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.

5 ACKNOWLEDGEMENTS

We thank the National Sciences and Engeneering Research Council (NSERC) of Canada, Mathematics of Information Technology and Complex Systems (MITACS) grant, and Life Technologies (Applied Biosystems) for funding.

REFERENCES

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., Vitulo, N., and Valle, G. (2009). Pass: a program to align short sequences. *Bioinformatics*, **25**(7), 967–968.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., and Frazer, K. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**(3), R32+.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, **10**(3), R25+.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, **25**(14), 1754–1760.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome research*, **18**(11), 1851–1858.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). Soap: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, **24**(5), 713–714.

Li, R., Yu, C., Li, Y., Lam, T.-W. W., Yiu, S.-M. M., Kristiansen, K., and Wang, J. (2009). Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics (Oxford, England)*, **25**(15), 1966–1967.

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B., and Li, M. (2008). Zoom! zillions of oligos mapped. *Bioinformatics*, **24**(21), 2431–2437.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P.-Y., and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, **23**(4), 452–456.

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M., and Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research*, **19**(9), 1527–1541.

CS	VARiD - AB map				VARiD - SHRiMP map				Corona			
data	Precision	Recall	TP	FP	Precision	Recall	TP	FP	Precision	Recall	TP	FP
CS-10x	88.5	17.3	69	9	76.7	34.7	138	42	82.3	16.3	65	14
CS-20x	85.1	31.7	126	22	74.6	51.0	203	69	79.6	33.4	133	34
CS-50x	81.1	47.5	189	44	80.0	67.3	268	67	80.2	57.0	227	56
CS-100x	84.0	59.5	237	45	77.9	73.4	292	83	81.8	69.8	278	62
CS-200x	83.0	64.8	258	53	79.9	77.1	307	77	80.1	74.6	297	74

Table 1. 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 run VARiD on various datasets aligned via the AB aligner, in the second we run it with SHRiMP mappings, and finally in the third we run the Corona Lite pipeline on the AB SOLiD mappings, for which it is optimized. In general the results show that variation detection is difficult even with high coverage of color-space, and the results are hard to compare among the pipelines - for example, VARiD with SHRiMP mappings tends to have lower precision, but higher recall. These results are improved significantly when adding just a small amount of 454 coverage in the combined VARiD platform, as seen in Table 3.

LS	VARiD - mosaik map				VARiD - SHRiMP map				gigaBayes			
data	Precision	Recall	TP	FP	Precision	Recall	TP	FP	Precision	Recall	TP	FP
LS-1x	39.2	10.8	38	59	62.1	11.5	41	25	80.4	11.3	45	11
LS-2.5x	67.4	31.5	124	60	73.1	31.0	122	45	70.2	34.9	139	59
LS-5x	63.0	50.0	199	117	74.0	49.2	196	69	64.0	59.8	238	134
LS-10x	74.5	68.3	272	93	76.4	67.6	269	83	59.2	68.8	274	189
LS-20x	67.7	82.2	327	156	70.3	83.2	331	140	55.8	64.3	256	203

Table 2. Results of running VARiD (Mosaik Alignments), VARiD (SHRiMP 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 can be argued to perform similarly at the lower (under 10x) coverage, while VARiD with Mosaik alignments runs slightly worse. However, in the higher coverages, VARiD performs better, with both higher precision as well as equal or better recall. For example, at 20x, VARiD with SHRiMP mappings has 70% precision to gigaBayes' 56%, and has 83% recall to gigaBayes' 64%.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., and Brudno, M. (2009). Shrimp: Accurate mapping of short color-space reads. *PLoS Comput Biol*, **5**(5), e1000386+.

Letter-space Coverage

Colour-space Coverage	**0x**	**1x**	**2.5x**	**5x**	**10x**
0x		62.1 11.5	73.1 31.0	74.0 49.2	76.4 67.6
10x	76.7 34.7	78.2 38.7	78.8 47.7	80.5 61.1	80.2 73.1
20x	74.6 51.0	76.9 47.7	81.8 54.3	82.4 66.1	81.2 79.4
50x	80.0 67.3	81.1 61.3	82.8 66.3	83.8 76.6	83.8 83.2
100x	77.9 73.4	80.1 70.9	82.8 73.6	84.7 80.7	84.0 88.2

Table 3. These numbers show the improvements we can obtain when combining reads from various platforms. Comparing at cost, for example, we can look at combining 50x of color-space data with 5x of 454 Roche data. Comparing to the equivalent cost of 454 Roche data at 10x in Table 2, We find that we are around 7% more precise and have 9% better recall rate in the combined run. Comparing to the cost equivalent of AB SOLiD Color-space at 100x, we obtain around a 6% better precision and 3% better recall. Another example can be found by looking at the CS-100x and LS-10x combination, and comparing with 200x of CS or 20x of LS in the previous Tables.