Rare variant detection in viral and bacterial populations

V.Y. Fofanov, M. Shin, D. Kim, H. Koshinsky

Abstract

Identification of variants within relatively homogeneous biological samples, such as single-species bacterial or viral populations, is important in many fields, including forensics, infectious disease research, biodefense, and biologics quality control. Advances in High Throughput Sequencing (HTS) allows sensitive detection of SNPs at previously unattainable frequencies – potentially down to ultra-rare variants present in < 0.1% frequency of the test population.

It is generally recognized that errors can and will be introduced at every step of the sequence data generation process, from DNA fragmentation and through to base assignment by the HTS machine. In addition to sequencing errors, reference sequence-based effects and mapping algorithm errors have two significant impacts on rare variant (SNP) detection accuracy: (1) some SNPs are more difficult to detect (loss of sensitivity), (2) some SNPs are more likely to be incorrectly detected (loss of specificity).

We have developed an exhaustive approach to identify and quantify **all** one-mismatch-away effects of reference genome interference and mapping algorithm associated errors. This approach was tested on *in silico* simulated viral and bacterial populations and on plasmid mixtures, with known SNP containing plasmids present in as low as 0.1% frequency. Multiple HTS platforms (IonTorrent, SOLiD, and HiSeq) and library preparation protocols were evaluated on the basis of accuracy of rare variant detection. The results suggest that the reference genome features, for a given n-mer length, have a significant impact on both the sensitivity (through silencing effects) and specificity (through ambiguous called variant) of variant detection. When proper correction procedures are applied, these errors can be significantly mitigated, making rare variant detected feasible – potentially even at ultra-rare variant frequencies.

Introduction / Sources of Error

Our Focus

- Rare variant (SNP) detection of mutations present in as low as 0.1% of single-species viral or bacterial cell population (1:1,000 mixture ratio)
- Whole-genome analysis - not constrained to preselected loci and capable of detecting previously unknown variants
- Focus on SNPs (including short indels), rather than on large insertions/deletions or genome rearrangements

Library preparation / Sequencing platform errors

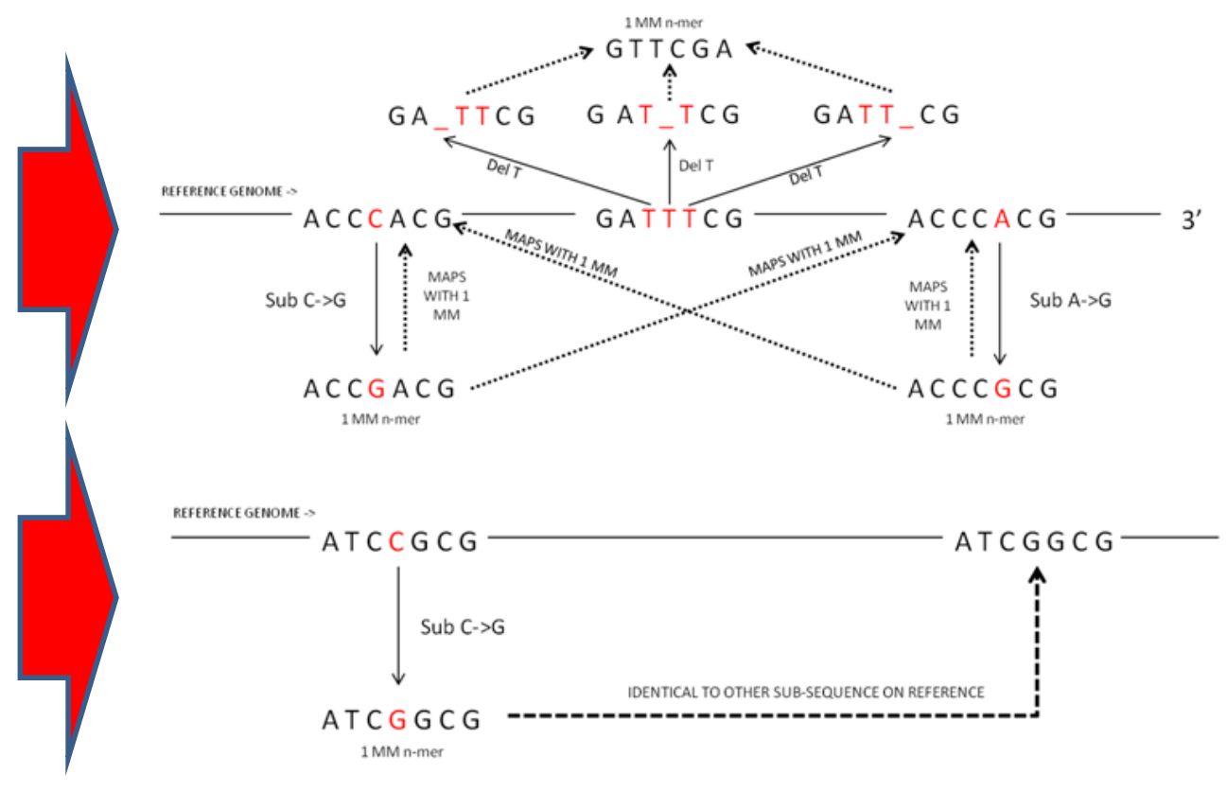
The amount, type, and location of sequencing errors within reads generated by HTS platforms have a significant impact in the sensitivity of rare variant detection and affect the confidence of the rare variant calls being made. Important factors include:

- Global error rates associated with platform / library protocol
- Spatial distribution of errors along the read
- Patterns of errors (e.g. homo-polymer extension errors)

Reference genome interference and associated errors

Features of reference genome may cause some variants to be ambiguous and particularly prone to false positive calls

…Other variants can be more difficult or simply impossible to detect



Mapping algorithm associated errors

Short reads (like those produced by HTS) mapping algorithms are not perfect. Assuming an alignment is possible, typical problems include:

- Un-mapped / missed reads
- Incorrectly mapped reads (no correct alignment is found)
- Ambiguously mapped reads (correct alignment is found, but 1 or more incorrect alignments are also reported)

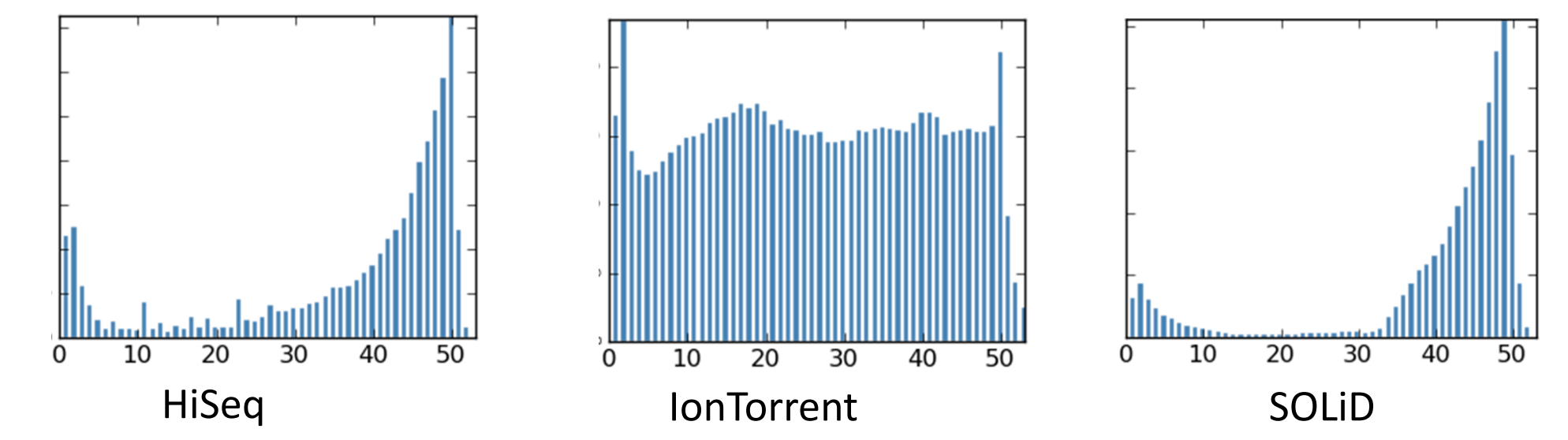
Rare Variant Detection Approach

Quantifying the effects of sequencing platform and library preparation

We have quantified the effect of several sequencing platforms (HiSeq, SOLiD, IonTorrent) and library preparation protocols (Nebulization, Bioruptor, Fragmentase, and Nextera) on the frequency, location, and type of sequencing errors introduced into a custom designed test plasmid. This was done to determine the optimal choice for platform/library (based on sequencing error profiles, genome coverage stability and throughput) and to calibrate the SNP annotation models for the type of errors that are expected to arise from each platform / library preparation combination.

Choice of sequencing platform affects spatial distribution of errors along the reads

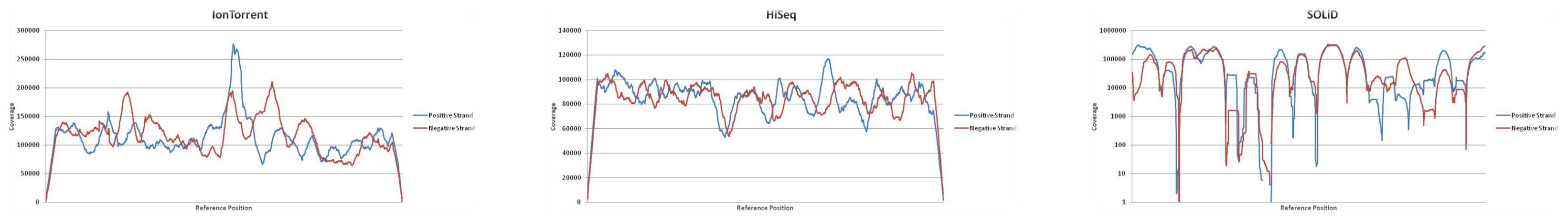
Spatial distribution of errors along the read for HiSeq, IonTorrent, and SOLiD platforms



Platform	Global Error Rate	Substitution Error Rate	Deletion Error Rate	Insertion Error Rate
HiSeq	0.928%	0.446%	0.157%	0.325%
SOLiD	0.650%	0.316%	0.166%	0.169%
IonTorrent	3.331%	0.159%	1.470%	1.703%

Global error rate is more influenced by sequencing platform than by library preparation protocol

Choice of sequencing platform and library preparation protocol affects genome coverage stability



Quantifying the effects of reference genome interference

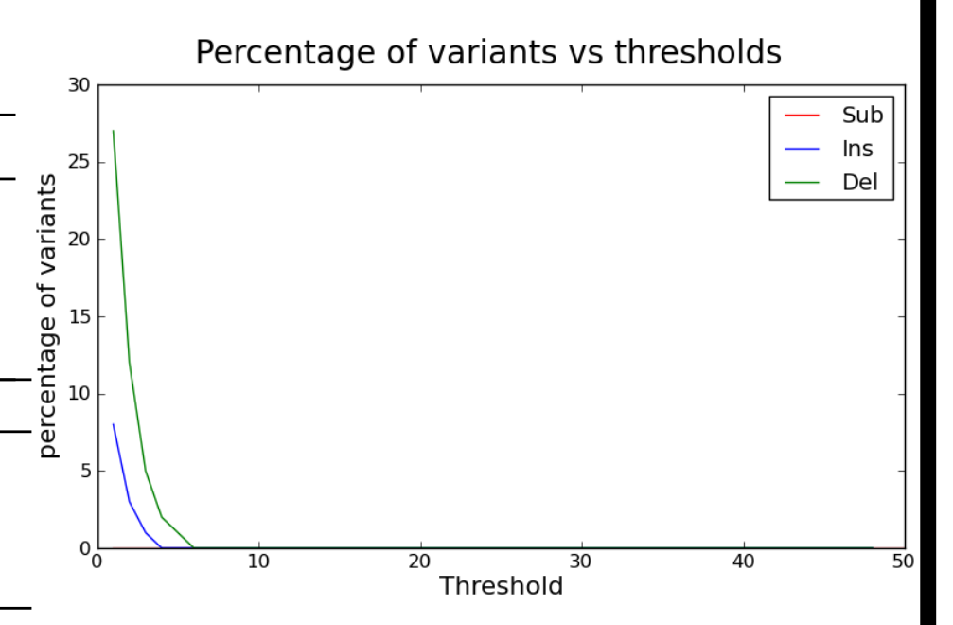
We have developed an automated pipeline that is capable of exhaustively simulating all possible one-mismatch-away reads of length *n* (n-mers) that can be generated from every subsequence of length n within the target reference genome (positive and negative strands simultaneously). Evaluating such n-mers allows us to quantify the degree of silencing and ambiguity for every possible variant (3 possible substitutions, 4 insertions, 1 deletion) on every position of a given reference genome sequence. The pipeline has been used to quantify the effects of reference genome interference in *B. anthracis Ames* for n-mer size of 50nt.

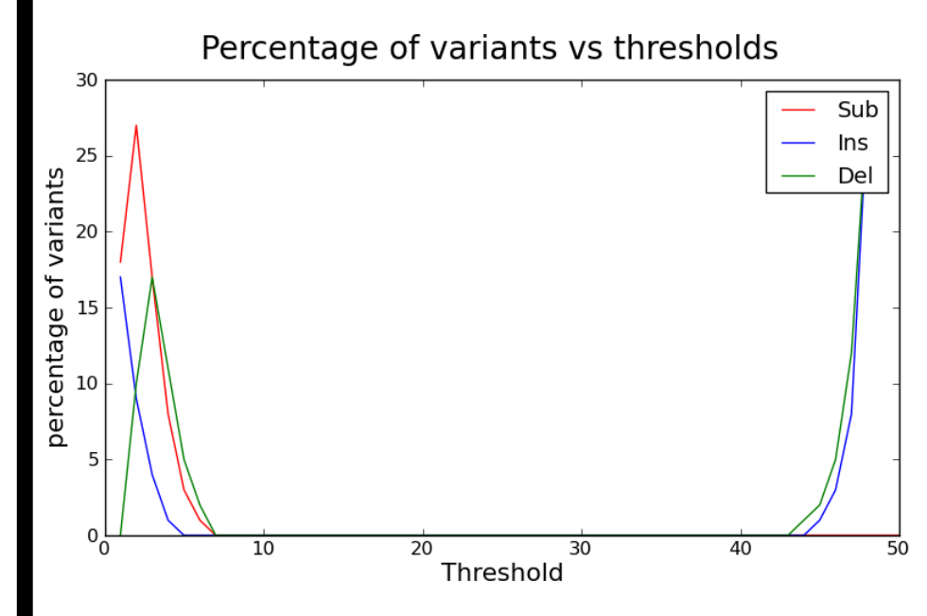
- Complete silencing (variant rendered completely undetectable) is rare
- Partial silencing (up to 20% loss in total reads capable of confirming a given variant) affects a significant fraction of variants (particularly insertions and deletions)

Number and percentage of variants affected by silencing

	Substitutions	Insertions	Deletions
Unaffected	15,680,402	17,944,436	2,598,381
Partially silenced (under 20%)	940	2,964,296	2,628,739
Completely Silent	181	29	61

	Substitutions	Insertions	Deletions
Unaffected	99.992%	85.823%	49.709%
Partially silenced (under 20%)	0.006%	14.177%	50.290%
Completely Silent	0.001%	0.000%	0.001%





Number and percentage of variants affected by ambiguous alignments

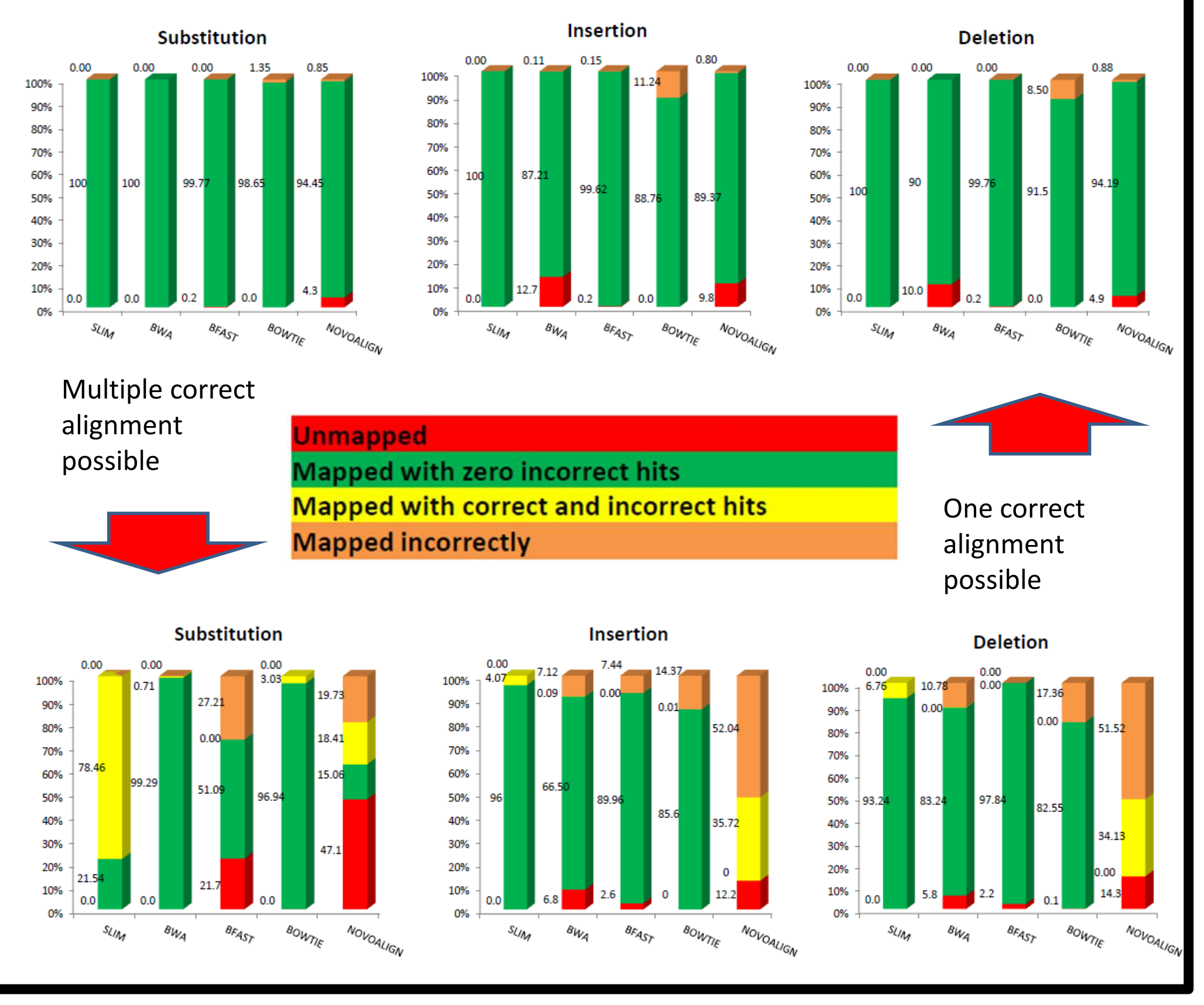
	Substitutions	Insertions	Deletions
Unaffected	3,484,090	4,782,989	61
Partially ambiguous (under 20%)	12,035,599	7,121,513	2,570,501
Completely ambiguous	144,005	6,026,714	1,465,145

	Substitutions	Insertions	Deletions
Unaffected	22.218%	22.876%	0.001%
Partially ambiguous (under 20%)	76.750%	34.060%	49.176%
Completely ambiguous	0.918%	28.824%	28.029%

- Both partial and complete ambiguity affects a significant fraction of possible variants (particularly for insertions and deletions)
- Repeatable regions, low complexity regions (e.g. ATATAT), homo-polymer are most common sources

Quantifying the effects of mapping algorithm errors

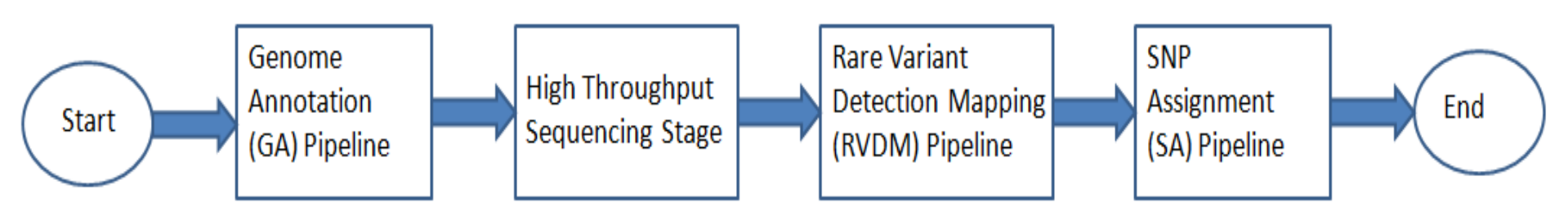
- The same automated pipeline was used to detect mapping errors
- All possible one-mismatch-away sequences that can be generated from the reference genome can be used to evaluating accuracy of the mapping algorithm
- Multiple mapping algorithms have been evaluated. No mapping algorithm is perfect (some are better with insertions and deletions, while others are better with substitutions)
- The probability of a mapping algorithm error can be annotated for every variant on every position of a reference genome



Results

Rare variant detection pipeline

The rare variant detection pipeline has been developed to automatically perform the annotation steps (identifying variants affected by reference genome interference and mapping algorithm errors):


Start → Genome Annotation (GA) Pipeline → High Throughput Sequencing Stage → Rare Variant Detection Mapping (RVDM) Pipeline → SNP Assignment (SA) Pipeline → End

The SNP Assignment module takes into account:

- Coverage fluctuations (adjusted for reference silencing effects)
- Mapping algorithm introduced errors
- Reference genome interference (false positives due to ambiguously aligned reads)
- Sequencing platform / library preparation associated errors and sequencing quality scores (if available)

Testing and calibration

The rare variant detection pipeline has been validated on populations of artificial sequences of lengths from 0.3 – 4.8Kbp with known rare variants introduced at population frequencies from 0.1% - 1%.

Sample	*Total variants detected*	*Total TP*	*# of FP*
HiSeq 1	19	12	7
HiSeq 2	18	12	6
HiSeq 3	19	11	8
HiSeq 4	18	11	7
HiSeq 5	20	13	7
HiSeq 6	14	11	3

On HiSeq (best performing platform with best performing library preparation protocol), all true positive SNPs present at 1% frequency were detected with 0 false positive calls. All true positive SNPs present at 0.1% frequency were detected with 3.9 false positives detected per 1Kb of reference genome.

Conclusions

Next Generation Sequencing technologies (such as HiSeq, SOLiD, or IonTorrent) are capable of cost effectively producing enough sequencing data to enable identification of rare variants present in as low as 0.1% frequency within a population.

We have developed a rare variant detection pipeline that takes into account the three major sources of errors contributing to false positive variant calls – sequencing platform / library protocol associated errors, features of reference genome (interference) that make some variants more difficult to detect and others prone to false positive calls, and mapping algorithm associated errors.

We have tested the pipeline on designed sequence populations with known rare variants at differing frequencies (consistent with single-species viral populations). Testing in bacterial populations is pending.

Support

This work was supported in part by a DHS SBIR Phase I award N10PC20199 and SBIR Phase II award D11PC20117

Contact Information

Viacheslav Fofanov, PhD, Director - Bioinformatics
vyfofanov@eurekagenomics.com

Didier Perez, CFO
didier@eurekagenomics.com