Maximum Likelihood Estimation of Incomplete Genomic Spectrum from HTS Data

Serghei Mangul[a], Irina Astrovskaya[a],Marius Nicolae[b], Bassam Tork[a], Ion Mandoiu[b] and Alex Zelikovsky[a]

[a]Department of Computer Science, Georgia State University, Atlanta, GA 30303
[b]Department of Computer Science & Engineering, University of Connecticut, Storrs, CT 06269
{serghei,iraa,btork,alexz}@cs.gsu.edu,
{man09004,ion}@engr.uconn.edu

Abstract. High-throughput sequencing makes possible to process samples containing multiple genomic sequences and then estimate their frequencies or even assemble them. The maximum likelihood estimation of frequencies of the sequences based on observed reads can be efficiently performed using expectation-maximization (EM) method assuming that we know sequences present in the sample. Frequently, such knowledge is incomplete, e.g., in RNA-seq not all isoforms are known and when sequencing viral quasispecies their sequences are unknown. We propose to enhance EM with a virtual string and incorporate it into frequency estimation tools for RNA-Seq and quasispecies sequencing. Our simulations show that EM enhanced with the virtual string estimates string frequencies more accurately than the original methods and that it can find the reads from missing quasispecies thus enabling their reconstruction.

Keywords: high-throughput sequencing, expectation maximization, viral quasispecies, RNA-Sequencing

1 Introduction

With the advent of high-throughput sequencing (HTS) technologies, it becomes possible to sequence samples containing multiple genomic sequences and then attempt to estimate their frequencies or even assemble them. In this paper we will consider two such HTS applications:

(i) RNA-seq, when the transcriptome (library of isoforms) is known but may be incomplete and expression of isoforms (or genes) is estimated by their frequencies in the sample and
(ii) viral quasispecies sequencing, when the reference sequence of the viral strain is known but the task is to find sequences of distinct quasispecies which are slightly different from the reference as well as to estimate their frequencies in the sample.

The maximum likelihood estimation of frequencies of the sequences (further referred as strings) can be efficiently performed using expectation-maximization (EM)

method for the viral quasispecies application(see [4, 10, 1]) and for RNA-seq [1](see [7, 8]). In brief, the input to EM consists of a panel, i.e., a bipartite graph in which one part correspond to the strings and another correspond to the reads. An edge connecting a read with a string expresses the possibility of the read to be emitted by the string with the probability associated with the edge. Given a panel and frequencies of the reads, EM can find maximum likelihood estimate of string frequencies.

Although in the both applications a certain knowledge about the sequences in the sample is available, such knowledge (recorded in the panel) is frequently incomplete. In case of RNA-seq, not all isoforms are already in the databases and in case of viruses, initially, no quasispecies sequences are known. In this paper, we propose a new method of enhancing EM that tries to estimate the incompleteness of the panel obtaining more accurate estimates of string frequencies and identifying reads that are more probable to be emitted by missed strings.

The method adds a *virtual string* to the panel and then iteratively changes the panel by assigning reads to the virtual string. The proposed enhanced method, so called Virtual String EM (VSEM), has been incorporated into IsoEM [8] and ViSpA [1]. Our simulations show that the VSEM-enhanced methods (IsoVSEM and ViSpA-VSEM) estimate string frequency more accurately than the original methods and that ViSpA-VSEM can find the reads from missing quasispecies thus enabling their reconstruction.

The rest of the paper is organized as follows. The next section describes VSEM. In Section 3 we describe the IsoVSEM and results of its experimental validation on transcriptome libraries. Section 4 describes the combination ViSpA-VSEM of ViSpA and VSEM. In Section 5, we analyze experimental results comparing ViSpA, ViSpA-VSEM and ShorAH [10] on the simulated reads with and without sequencing errors.

2 Virtual String Expectation Maximization

In this section we first formally define the panel and briefly describe EM method. Then we show how to estimate the quality of the model. Finally we describe the VSEM method enhancing EM with the virtual string.

The input data for EM method consists of a *panel*, i.e., a bipartite graph $G = \{S \bigcup R, E\}$ such that each string is represented as a vertex $s \in S$, and each read is represented as a vertex $r \in R$. With each vertex $s \in S$, we associate unknown frequency f_s of the string. And with each vertex $r \in R$, we associate observed read frequency o_r. Then for each pair s_i, r_j, we add an edge (s_i, r_j) weighted by probability of string s_i to emit read r_j with m genotyping errors:

$$h_{s_i,r_j} = \binom{l}{m} (1-\epsilon)^{l-m} \epsilon^m,$$

where l is length of read sequence, and ϵ is the genotyping error rate.

Regardless of initial conditions EM algorithm always converge to maximum likelihood solution (see [3]).The algorithm starts with the set of N strings. After uniform initialization of frequencies $f_s, s \in S$, the algorithm repeatedly performs the next two steps until convergence:

[1] Note that frequency estimation based on previous approaches is less accurate (see e.g. [9]).

– E-step: Compute the expected number $n(j)$ of reads that come from string i under the assumption that string frequencies $f(j)$ are correct, based on weights $h_{s_i,j}$
– M-step: For each i, set the new value of f_s to to the portion of reads being originated by string s among all observed reads in the sample

In order to decide if the panel is incomplete we need to measure how well maximum likelihood model explains the reads. We suggest to measure the model quality by the deviation between expected and observed read frequencies as follows:

$$D = \frac{\sum_j |o_j - e_j|}{|R|},$$

where $|R|$ is number of reads, o_j is the observed read frequency of the read r_j and e_j is the expected read frequencies of the read r_j calculated as follows:

$$e_j = \sum_i \frac{h_{s_i,j}}{\sum_l h_{s_i,l}} f_i^{ML} \tag{1}$$

where $h_{s_i,j}$ is weighted match based on mapping of read r_j to string s_i and f_j^{ML} is the maximum-likelihood frequency of the string s_i.

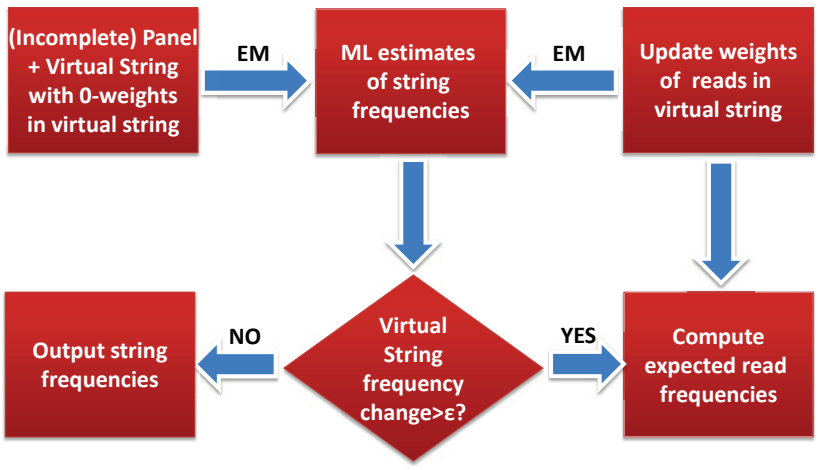

Fig. 1. Flowchart for VSEM.

The main idea of the VSEM algorithm (see Algorithm 1) is to add to set of candidate strings a virtual string which emits reads that do not fit well to existing sequences. The flowchart of VSEM is on Fig. 1. Initially, all reads are connected to the virtual string with weight $h_{s_i,j} = 0$. The first iteration finds the ML frequency estimations of candidates strings, ML frequency estimations of virtual string will be equal to 0,

Algorithm 1 VSEM algorithm

add virtual string vs to the set of candidate strings
initialize weights $h_{vs,j} = 0$
while $\Delta vs > \epsilon$ **do**
  calculate f_j^{ML} by EM algorithm
  $e_j = \sum_i \frac{h_{s_i,j}}{\sum_l h_{s_i,l}} f_i^{ML}$
  $D = \frac{\sum_j |o_j - e_j|}{|R|}$
  $\delta = o_j - e_j$
  if $\delta > 0$ **then**
    $h_{vs,j} += \delta$
  else
    $h_{vs,j} = \max\{0, h_{vs,j} + \delta\}$
  end if
end while

since all edges between virtual string and reads $h_{vs,j} = 0$. Then these estimation are used to compute expected frequency of the reads according to (1). If the expected read frequency is less than the observed one (under-estimated), then the lack of the read expression is added to the weight of the read connection to the virtual string. For over-estimated reads, the excess of read expression is subtracted from the corresponding weight (but keeping it non-negative). The iterations are continued while the virtual string frequency is decreasing by more than ϵ.

Enhancing of ViSpA. Resulted edge weight between virtual string and each read can be interpreted as the probability of the read to be emitted by missing strings. VSEM transmits to ViSpA assembler the rounded read weights and during assembling of additional candidate quasispecies the preference is given to reads which are likely emitted by missing strings.

Enhancing of IsoEM. The IsoEM incorporates the virtual string and the resulting isoform frequency estimations are improved. Based on the frequency of virtual string it is possible to decide if the panel is likely to be incomplete, the total frequency of missing strings is estimated by frequency of virtual string.

3 Experimental Validation of IsoVSEM on RNA-seq data

IsoEM is a novel expectation-maximization algorithm for inference of alternative splicing isoform frequencies from high-throughput transcriptome sequencing (RNA-Seq) data proposed in [8]. IsoEM takes advantage of base quality scores, strand information and exploits disambiguation information provided by the distribution of insert sizes generated during sequencing library preparation. In the bipartite graph consisting of isoforms and reads an edge from an isoform to a read represents possibility that a read is emitted by the isoform. It is noted [8] that EM can run in parallel for each connected component of this bipartite graph. We enhance IsoEM algorithm by adding virtual string to each connected component. The resulted algorithm IsoVSEM in the nested loop applies IsoEM instead of EM (see Algorithm 1). Since isoforms have different length we

estimate missing isoforms by volume defined as frequency of isoform multiplied by its length.

Our validation of IsoVSEM includes two experiments over human RNA-seq data. Below we describe the transcriptome data and read simulation and then give the settings for the each experiment and analyze the obtained experimental results.

Data sets. IsoVSEM was tested on human RNA-Seq data. The human genome data(hg18, NCBI build 36) was downloaded from UCSC and CCDS together with the coordinates of the isoforms in the KnownGenes table. The UCSC database contains 66803 isoforms from 19372 genes, and CCDS database contains 20829 isoforms from 17373 genes. Genes were defined as clusters of known isoforms defined by the GNFAtlas2 table such that CCDS data set can be identified with the subset of UCSC data set.

30M single error-free reads of length 25 were randomly generated by sampling fragments of isoforms from UCSC data set. Each isoform was assigned a true frequency based on the abundance reported for the corresponding gene in the first human tissue of the GNFAtlas2 table, and a probability distribution over the isoforms inside a gene cluster [8]. We simulate datasets with geometric (p=0.5) distributions for the isoforms in each gene.

Expression range		0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
MPE	Full panel	0.0	61.7	22.0	8.0	3.2	2.1	10.3
	Incomplete	0.0	59.3	41.3	24.8	19.7	5.9	33.7
	Incomplete + VS	0.0	47.2	33.1	20.7	16.4	8.5	26.9
$EF_{.15}$	Full panel	0.0	81.9	61.3	28.7	7.5	8.5	38.8
	Incomplete	0.0	81.7	72.4	61.4	56.7	42.1	67.6
	Incomplete+VS	0.0	77.2	68.2	57.6	53.0	36.8	63.6

Table 1. Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for isoform expression levels in Experiment 1.

Experiment 1: Comparison between IsoEm and IsoVSEM on reduced transcriptome data. We assumed that in every gene 25% of isoforms is missing. In order to create such an instance we assign to isoforms inside the gene geometric distribution(p=0.5), assuming a priori that number of isoforms inside the gene is less or equal to 3. This way we removed isoform with frequency 0.25. As a result 11339 genes were filtered out, number of isoforms was reduced to 24099. Note that in our data set missing isoforms do not have unique exon junctions that can emit reads indicating that certain isoforms are missing.

We first check how well IsoVSEM can estimate the volume of missing strings. Although the frequencies of all missing strings (isoforms) is the same (25%), the volumes significantly differ because they have different length. Therefore, the quality can be measured by correlation between actual missing volumes and predicted missing volumes which are volumes of virtual strings. In this experiment it is 61% which is sufficiently high to give an idea which genes are missing isoforms in the database.

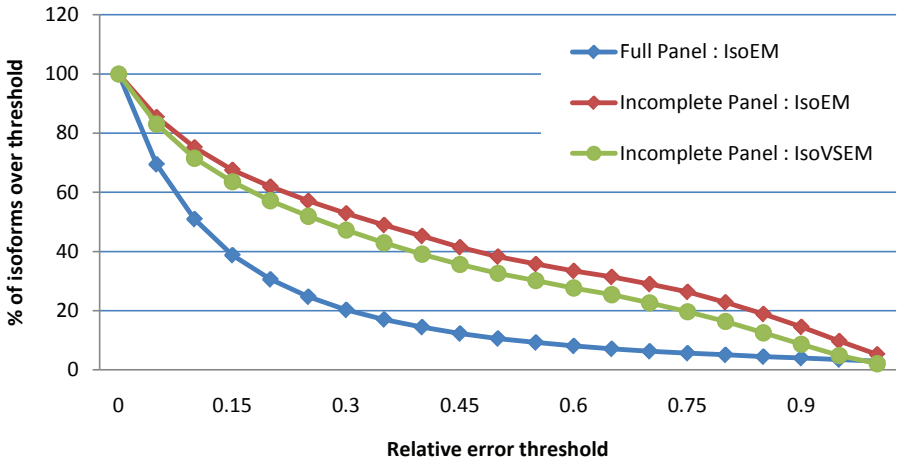

Fig. 2. Error fraction at different thresholds for isoform expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression. Blue line correspond to IsoEM with the full panel, read line is IsoEM with the incomplete panel, and green line is IsoVSEM.

Table 1 reports the median percent error (MPE) and .15 error fraction $EF_{.15}$ for isoform expression levels inferred from 30M reads of length 25, computed over groups of isoforms with various expression levels. MPE is the median relative error of isoform frequency estimation and the error fraction with threshold t, denoted EF_t, is defined as the percentage of isoforms with relative error greater or equal to t.

Figure 2 gives the error fraction at different thresholds ranging between 0 and 1. Clearly the best performance is achieved when the the isoform library is full, using virtual string explains accuracy gain of IsoVSEM over IsoEM. IsoVSEM achieves better accuracy in the case when the panel is incomplete. Performance of IsoEm and IsoVSEM for the full panel is the same.

Expression range		0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
MPE	Full panel	0.0	100	22.7	7.3	3.5	2.5	11.8
	Incomplete	0.0	100	45.5	29.4	28.5	28.7	31.8
	Incomplete + VS	0.0	100	43.2	27.09	25.68	14.34	29.61
$EF_{.15}$	Full panel	5.1	91.2	62.8	29.3	15.8	7.6	45.5
	Incomplete	18.6	95.6	85.6	83.3	89.2	86.7	80.0
	Incomplete+VS	17.6	91.8	81.3	77.9	80.3	75.5	75.2

Table 2. Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for isoform expression levels in Experiment 2.

Experiment 2: Comparison between IsoEm and IsoVSEM on the CCDS panel. In this experiment UCSC database represents the full set of isoforms and CCDS represents the incomplete panel. Reads were generated from UCSC library of isoforms, while only frequencies of known isoforms from CCDS database were estimated. In contrast to Experiment 1, we do not control the frequency of missing isoforms (i.e., isoforms from UCSC which are absent in CCDS). Therefore, one cannot expect as good improvements as in Experiment 1.

Table 2 reports the median percent error (MPE) and .15 error fraction $\mathrm{EF}_{.15}$ for isoform expression levels inferred from 30M reads of length 25, computed over groups of isoforms with various expression levels. We do not report the number of isoforms since they are different for UCSC and CCDS panels. Anyway, one can see a reasonable improvement in frequency estimation of IsoVSEM over IsoEM.

4 VSEM Enhancement of ViSpA

In this section we first give high-level description of ViSpA [1], a recent viral spectrum assembling tool for inferring viral quasispecies sequences and their frequencies from pyrosequencing shotgun reads. Then we describe the flowchart of the combining tool ViSpA-VSEM and required modifications to ViSpA.

ViSpA: Viral Spectrum Assembly. First, ViSpA aligns the reads to the consensus genome sequence using SEGEMEHL [6] software correcting obvious sequencing errors and removes subreads (reads that are completely covered by larger reads). Then it builds a read graph with vertices representing remaining reads (superreads) and edges representing overlaps between them. In this graph, each path from the leftmost vertex to the rightmost vertex corresponds to a possible candidate quasispecies sequence. For each edge e, ViSpA computes probability $p(e)$ to connect two reads from the same quasispecies. Then ViSpA assigns cost $-\log(p(e)) = \log(1/p(e))$ to each edge e, making the minimum-cost paths more probable to represent quasispecies sequences. Next, a set of candidate paths consisting of the max-bandwidth paths (paths minimizing maximum edge cost) through each vertex is created and refined so that only distinct sequences remain. The maximum-likelihood estimates of frequencies are calculated by EM algorithm which takes in account all reads in the sample. Finally, ViSpA reports most frequent candidate sequences and their frequencies as inferred viral quasispecies spectrum.

Combining ViSpA with VSEM. Knowing which reads in a sample are likely to be produced from missing (unknown) quasispecies sequences may allow to expand ViSpA's candidate set. Additionally, we can improve ViSpA's estimates for quasispecies frequencies by taking in account incompleteness of the panel.

Figure 3 illustrates the proposed workflow between (modified) ViSpA and VSEM. At each iteration, ViSpA gets a set of aligned reads and their 0/1-weights estimating probability to be emitted by unknown strings (candidates). Initially, all reads have weight zero and ViSpA works as described above except the maximum likelihood estimates for candidate quasispecies sequences (strings) are calculated by VSEM. On all other iterations, VSEM panel includes not only newly assembled sequences but also

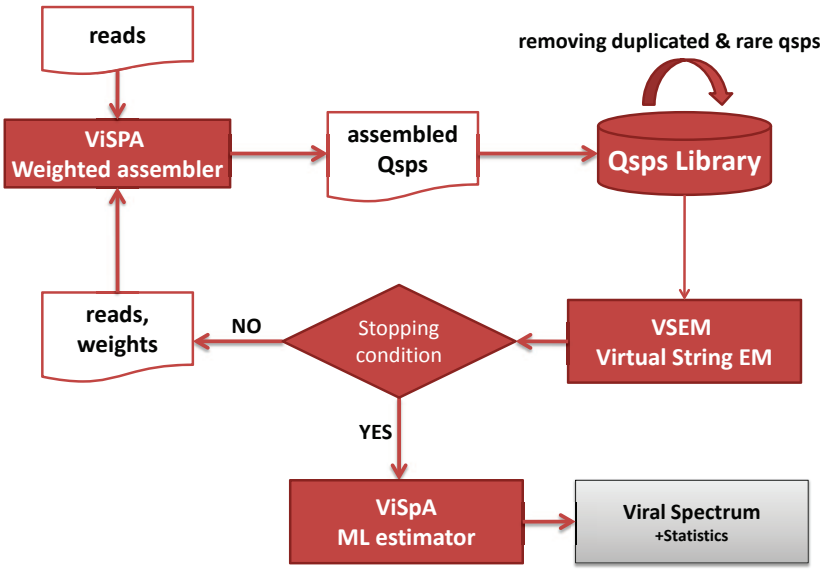

Fig. 3. Flowchart of ViSpA-VSEM

candidate sequences stored from the previous iterations. All sequences with negligible EM frequency are filtered out from the cumulative set of sequences, called quasispecies library. Once frequencies distribution is obtained, VSEM assigns 0/1-weight for each read: 0 corresponds to high probability to be emitted by a missing sequences and 1, otherwise. Finally, if the virtual string has high EM frequency and we expand our quasispecies library with respect to previous iteration, we feed reads and their weights back to ViSpA, and all process is repeated. At the end, we report sequences from quasispecies library and their frequencies as reconstructed viral quasispecies spectrum.

The modifications of ViSpA in ViSpA-VSEM include (1) superread selection (a weight-1 superread is removed if it shares a subread with a 0-weight superread), (2) edge cost computation which takes in account vertex weights:

$$cost'(e) = cost(e) + 0.5 \cdot L \cdot (w(u) + w(v)),$$

where $cost(e)$ is the original cost of e, L is the read length, and $w(u)$, $w(v)$ are the 0/1 weights assigned by VSEM to read v.

5 Experimental Validation of ViSpA-VSEM on Simulated Data

Data Sets. We simulate reads from 1739-bp long fragment from the E1E2 region of 44 HCV sequences [5]. Each population was created by randomly selecting either 10 or 40 sequences among these HCV variants and assigning frequencies following either

(1) uniform (all sequences have the same frequency), or (2) skewed uniform (a single sequence has high frequency; all other sequences have uniformly low frequency), or (3) geometric (the i^{th} sequence is a constant percentage more frequent that the $(i+1)^{th}$ sequence) distributions.

First, we simulate error-free reads without indels with respect to the reference sequence. The length of a read follows normal distribution with variance 400, and starting position follows the uniform distribution. This simplified model of reads generation has two parameters: number of the reads that varies from 20K up to 100K and the averaged read length that varies from 100bp up to 500bp.

Then we simulate 454 pyrosequencing reads from the 10 random quasispecies (under geometric distribution) out of 44 HCV sequences [5] using FlowSim [2]. The generated dataset contains 39,131 reads with length varying from 50bp up to 550bp and average length equaled to 322bp. Each position (except the end) is covered by at least 4000 reads. 99.96% of aligned reads has at least one indel with respect to the reference: 99.97% of deletions and 99.6% of insertions are 1bp long. Only 1.1% of aligned reads have unknown base(s).

Frequency estimation quality. We evaluate predicted frequencies by the following statistics.

- Kullback-Leibler divergence

$$RE = \sum_{i \in I} p_i \log \frac{p_i}{q_i},$$

  where $P = \{p_i\}$ and $Q = \{q_i\}$ are true distribution and its approximation, and $I = \{i | p_i > 0,\, q_i > 0\}$ are real sequences among assembled candidate sequences,
- correlation between real and predicted frequencies,
- average prediction error:

$$\overline{err} = \frac{\sum_{i \in I} |p_i - q_i|}{|I|}.$$

Detection of panel incompleteness. We have checked how well VSEM can detect incompleteness of the panel in the following experiment. We have repeatedly (for different simulated frequency distribution for 10 quasispecies strings) deleted from the full panel each string (one at a time) and record the resulted frequency of the virtual string. If no string has been deleted, then virtual string has always stopped growing at frequency less than 10^{-6}, and if the frequency of the deleted string has been at least 1%, then the resulted virtual string frequency has grown to at least .5%. Thus VSEM can reliably detect incomplete panel if missing strings have total frequency at least 1%.

Improving quasispecies frequencies estimation using VSEM. Fig. 3 show our experimental results on simulated error-free reads generated from 40 quasispecies. The correlation is slightly improved for cases when the portion of missing strings is small and increases to as much as 15% when bigger portion of strings is missing.

The results on reads generated by Flowsim [2] and corrected by ShorAH are very similar to the results on error free reads.

	r.l./n.r	% of missing strings											
		$< 10\%$		10%-20%		20%-30%		30%-40%		40%-50%		$> 50\%$	
		r	err	r	err	r	err	r	err	r	err	r	err
ViSpA	100/20K	90.2	4.5	91.0	6.8	75.4	5.1	68.6	1.6	40.8	2.3	39.8	10.4
ViSpA-VSEM	100/20K	91.6	2.3	92.8	4.4	76.5	4.1	70.5	1.4	54.2	2.0	50.8	7.4
ViSpA	300/20K	95.7	3.8	93.2	10.2	89.8	1.0	66.7	1.5	62.1	2.1	46.8	9.7
ViSpA-VSEM	300/20K	95.4	1.7	95.8	1.1	96.9	0.6	85.7	0.9	88.0	0.9	60.4	2.6
ViSpA	100/100K	95.2	4.5	93.9	9.1	84.8	1.4	74.2	1.8	74.5	2.3	73.4	9.9
ViSpA-VSEM	100/100K	97.8	2.6	95.6	3.0	86.3	1.3	79.8	1.7	79.0	2.1	74.2	8.8
ViSpA	300/100K	96.2	3.9	88.6	12.4	88.9	1.0	85.1	1.4	75.1	2.3	49.5	10.5
ViSpA-VSEM	300/100K	96.2	2.0	92.8	0.9	93.7	0.7	90.2	1.2	84.4	1.7	67.1	4.8

Table 3. Correlation (r) and average prediction error (err) between real quasispecies frequencies and estimated quasispecies frequencies for EM vs VSEM. r.l/n.r denote the read length / number of reads.

Detection of reads emitted by missing strings. The output of VSEM besides estimated frequency of the virtual string also contain the weights of edges connecting reads to the virtual string. These weights can be interpreted as probabilities of reads to be emitted by the missing strings. In our experiments we have repeatedly measure the correlation between the edge weights and the spectrum of reads emitted by missing strings which has always exceeded 65%.

ViSpA versus ViSpA-VSEM. We compare quality of assembling and frequency estimation for both methods. Quality of assembling is measured by sensitivity (portion of the assembled real sequences among all real quasispecies) and its positive predictive value (portion of the real sequences among all assembled) in cross-validation tests.

Error-free reads. Previously [1], we demonstrate that ViSpA outperforms SHORAH in assembling haplotypes on error-free reads. ViSpA-VSEM can further improve predictive power and frequency estimation of ViSpA (see Table 4). On average, it infers additional two (in case of geometric distribution) to four sequences (in case of uniform and skewed uniform distributions). Taking into account unknown quasispecies sequences allows ViSpA-VSEM to estimate frequencies more accurately (average error is decreased 2.5 times for geometric distribution and more than 5 times for skewed uniform and uniform distributions). Since relative entropy and correlation coefficient r are measured only on the correctly inferred quasispecies sequences and are not adjusted with respect to the number of all quasispecies sequences in a sample, increasing relative entropy and decreasing of correlation coefficient r are not correlated with loss of predictive power. For example, predictive power is improved by obtaining additional real quasispecies in the case of geometric distribution whereas correlation coefficient becomes smaller.

Reads with simulated genotyping errors. It has been shown that ViSpA outperforms SHORAH if sequencing errors are initially corrected (see [1]). So in our experiments, we compare ViSpA and ViSpA-VSEM only on ShoRAH-corrected reads (see Fig. 5). The table reports the difference between 10 most frequent assemblies obtained by ViSpA and 10 most frequent assemblies obtained after two iterations of ViSpA-VSEM.

	ViSpA					ViSpA-VSEM					
Distribution	PPV	SE	RE	r	err	PPV	SE	RE	r	err	Gain
Geometric	0.767	0.5	-0.0099	0.954	7.36	0.5905	0.73	0.0276	0.9094	2.91	2.3
Skewed	0.733	0.4	-0.0196	0.6725	13.01	0.701	0.77	0.0085	0.9665	2.5	4
Uniform	0.733	0.4	-0.0191	0.716	12.76	0.645	0.73	0.0108	0.9762	2.34	3.7

Table 4. Comparison between ViSpA and ViSpA-VSEM. Experiments are run on 100K reads from 10 quasispecies with average read length equaled to 300. The table reports PPV, sensitivity(SE), relative entropy (RE), correlation between real and predicted frequencies (r), and averaged prediction error ($\overline{err}$)(reported in %). "Gain" column reports averaged number of additionally inferred real quasispecies sequences after 4 iterations (on average) for skewed distribution, 5 iterations (on average) for geometric distribution and 13 iterations (on average) for uniform distribution.

ViSpA-VSEM can additionally infer a real quasispecies without allowing any mismatches between sequences($k = 0$). Again, the frequency estimation is more accurate since ViSpA-VSEM EM takes into account missing quasispecies which is confirmed by the drop of the average prediction error.

	ViSpA					ViSpA-VSEM					
#mismatches	PPV	SE	RE	r	err	PPV	SE	RE	r	err	Gain
k = 0	0.5	0.5	0.0720	0.9860	9.98	0.5455	0.6	0.0494	0.9741	7.54	1
k = 2	0.6	0.6	0.0668	0.9860	9.16	0.6364	0.7	0.0434	0.9680	6.67	1
k = 6	0.7	0.7	0.0577	0.9856	7.95	0.7273	0.8	0.0369	0.9463	6.20	1
k =7	0.8	0.8	0.0525	0.9866	7.26	0.8182	0.9	0.0335	0.9479	5.65	1

Table 5. Comparison between ViSpA and ViSpA-VSEM on their 10 most frequent assemblies. Experimental results are run on 100K reads from 10 quasispecies with average read length equaled to 300. The quasispecies sequence is considered found if one of candidate sequences matches it exactly ($k = 0$) or with at most k (2, 6 or 7) mismatches. The table reports PPV, sensitivity(SE), relative entropy (RE), correlation between real and predicted frequencies (r), and averaged prediction error ($\overline{err}$)(reported in %). "Gain" column reports averaged number of additionally inferred real quasispecies sequences after 2 iterations.

6 Conclusions and Future Works

In this paper, we propose VSEM, a novel modification of EM algorithm which allows to estimate the frequencies of multiple genomic sequences present in a sample sequenced with HTS technology. VSEM is aimed to improve the maximum likelihood frequency estimations of assembled sequences and identify reads that belong to unassembled sequences. We have applied VSEM to enhance two tools: IsoEM (for inferring isofrom expression from RNA-seq data) and ViSpA (for inferring viral quasispecies spectrum from pyrosequencing shotgun reads). Our experimental study shows that VSEM-enhanced

tools significantly improve their performance: IsoVSEM has better accuracy in estimation isoform frequencies and ViSpA-VSEM can infer more quasispecies sequences and better estimate their frequencies. Our results show potential of VSEM to improve other metagenomics tools.

7 Acknowledgements

SM, IA and BT were supported in part by Molecular Basis of Disease Fellowship, Georgia State University. MN and IIM were supported in part by NSF awards IIS-0546457, IIS-0916948, and DBI-0543365. SM, IA, BT and AZ were supported in part by NSF award IIS-0916401. SM, IA, MN, BT, IM and AZ were supported in part by Agriculture and Food Research Initiative Competitive Grant no. 2011-67016-30331 from the USDA National Institute of Food and Agriculture.

References

1. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Mandoiu, P. Balfe, and A. Zelikovsky. Inferring viral spectrum from 454 pyrosequencing reads. *BMC Bioinformatics*, (to appear) url=http://dna.engr.uconn.edu/bibtexmngr/upload/Aal.11a.pdf.
2. S. Balser, K. Malde, A. Lanzen, A. Sharma, and I. Jonassen. Characteristics of 454 pyrosequencing data–enabling realistic simulation with flowsim. *Bioinformatics*, 26:i420–5, 2010.
3. N. Zaitlen, B. Pasaniuc and E. Halperin. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *Journal of Computational Biology*, 18(3): 459-468, March 2011.
4. N. Eriksson, L. Pachter, Y. Mitsuya, S.Y. Rhee, and C. Wang et al. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4:e1000074, 2008.
5. T. Von Hahn, J.C. Yoon, H. Alter, C.M. Rice, B. Rehermann, P. Balfe, and J.A. Mckeating. Hepatitis c virus continuously escapes from neutralizing antibody and t-cell responses during chronic infection in vivo. *Gastroenterology*, 132:667–678, 2007.
6. S. Hoffmann, C. Otto, S. Kurtz, C.M. Sharma, P. Khaitovich, J. Vogel, P.F. Stadler, and J. Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 09 2009.
7. B. Li, V. Ruotti, R.M. Stewart, J.A. Thomson, and C.N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.
8. M. Nicolae, S. Mangul, I.I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-seq data. *Algorithms for Molecular Biology*, 6:9, 2011.
9. A. Mortazavi, B.A.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 2008.
10. O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *Journal of computational biology : a journal of computational molecular cell biology*, 17(3):417–428, March 2010.