Reconstruction of Viral Population Structure from Next-Generation Sequencing Data Using Multicommodity Flows

Pavel Skums[1*§], Nicholas Mancuso[2*], Alexander Artyomenko[2*], Bassam Tork[2], Ion Mandoiu[3], Yury Khudyakov[1] and Alex Zelikovsky[2]

[1] Laboratory of Molecular Epidemiology and Bioinformatics, Division of Viral Hepatitis, Centers for Disease Control and Prevention, 1600 Clifton Road NE, 30333 Atlanta, GA, USA

[2] Department of Computer Science, Georgia State University, 34 Peachtree str., 30303, Atlanta, GA, USA

[3] Department of Computer Science and Engineering, University of Connecticut, 06269, Storrs, CT, USA

*These authors contributed equally to this work

[§]Corresponding author

Email addresses:

PS: kki8@cdc.gov

AA: aartyomenko@cs.gsu.edu

NM: nmancuso @cs.gsu.edu

BT: btork @cs.gsu.edu

IM: ion@engr.uconn.edu

AZ:alexz@cs.gsu.edu

YK:yek0@cdc.gov

Abstract

Background

Highly mutable RNA viruses exist in infected hosts as heterogeneous populations of genetically close variants known as quasispecies. Next-generation sequencing (NGS) allows for analysing a large number of viral sequences from infected patients, presenting a novel opportunity for studying the structure of a viral population and understanding virus evolution, drug resistance and immune escape. Accurate reconstruction of genetic composition of intra-host viral populations involves assembling the NGS short reads into whole-genome sequences and estimating frequencies of individual viral variants. Although a few approaches were developed for this task, accurate reconstruction of quasispecies populations remains greatly unresolved.

Results

Two new methods, AmpMCF and ShotMCF, for reconstruction of the whole-genome intra-host viral variants and estimation of their frequencies were developed, based on Multicommodity Flows (MCFs). AmpMCF was designed for NGS reads obtained from individual PCR amplicons and ShotMCF for NGS shotgun reads. While AmpMCF, based on covering formulation, identifies a minimal set of quasispecies explaining all observed reads, ShotMCS, based on packing formulation, engages the maximal number of reads to generate the most probable set of quasispecies. Both methods were evaluated on simulated data in comparison to Maximum Bandwidth and ViSpA, previously developed state-of-the-art algorithms for estimating quasispecies spectra from the NGS amplicon and shotgun reads, respectively. Both algorithms were accurate in estimation of quasispecies frequencies, especially from large datasets.

Conclusions

The problem of viral population reconstruction from amplicon or shotgun NGS reads was solved using the MCF formulation. The two methods, ShotMCF and AmpMCF, developed here afford accurate reconstruction of the structure of intra-host viral population from NGS reads. The implementations of the algorithms are available at http://alan.cs.gsu.edu/vira.html (AmpMCF) and http://alan.cs.gsu.edu/NGS/?q=content/shotmcf (ShotMCF)

Background

RNA-dependent RNA-polymerases of RNA viruses are error prone and frequently generate mutations, accumulation of which results in a diverse intra-host viral population of closely related genetic variants [1,3], commonly termed quasispecies by virologists.

The advent of Next-Generation Sequencing (NGS) presented invaluable opportunity for the in-depth evaluation of viral quasispecies and understanding the structure of viral intra-host populations in unprecedented detail [18, 19]. The application of NGS to clinical and public health settings offers prospects for significant improvement in controlling drug resistance [13] and development of novel therapeutics and vaccines [17]. One of the major advantages of NGS is in generating sequence data at a scale that allows not only analysis of intra-host viral variants from a single amplicon or recovery of the consensus full-length genomic sequence but also reconstruction of the population of full-genome quasispecies from an infected host.

The problem of reconstruction of a structure of viral population formulated as *quasispecies spectrum reconstruction problem* was recently addressed in several studies [6, 7, 8, 9, 15]. Given a collection of the shotgun or amplicon NGS reads

generated from a sample of the viral population, the algorithms reconstruct a set of quasispecies and their relative frequencies. All published algorithms are based on generating graphs of read overlaps and use minimum-cost flows, probabilistic methods, shortest paths, or maximum bandwidth to reconstruct a set of quasispecies from the graphs [6, 7, 8, 9, 15]. The accuracy of reconstruction is affected by the heterogeneity of intra-host viral population. The abundance of conserved genomic regions that extend beyond the maximal read length significantly restricts the full-genome quasispecies assembly. Indeed, even short conserved regions at the overlaps of reads significantly increase ambiguity of quasispecies reconstruction.

Most algorithms for the quasispecies spectrum reconstruction implicitly assume that sequence data were obtained using a shotgun experiment. Although the shotgun method is frequently used for reconstruction of long sequences and produces less distortion in frequency of quasispecies than the amplicon-based approach, the available NGS error correction algorithms are most efficient when applied to amplicon-based data [4,12]. Additionally, although most quasispecies spectrum reconstruction algorithms are technically applicable to both types of data, the amplicon-based approaches allow for a greater control over the distribution of reads across the entire sequence of interest, resulting in a more accurate estimation of the structure of viral population [8,9].

In this paper, we consider two methods, AmpMCF and ShotMCF, for reconstruction of the genetic structure of intra-host viral population using either amplicon or shotgun NGS reads, respectively. Both methods are based on the application of MultiCommodity Flow problem (MCF) [21].

Methods

MCF is a classical optimization problem that searches for k flows for k source-sink pairs (s_i, t_i) in a network N in order to either minimize the total cost of flows or maximize the total flow of subjects to match the capacity and demand constraints.

Quasispecies reconstruction can be formulated as an optimization problem in two ways: (1) identification of the most probable set of quasispecies formed by the largest subset of reads from the data, referred to as packing formulation; and (2) identification of a minimal set of quasispecies explaining all observed reads, referred to as covering formulation. These two formulations, when applied to MCFs, were developed into path packing and path covering algorithms of ShotMCH and AmpMCF, respectively.

AmpMCF algorithm

We consider an amplicon A as a multiset of reads such that each read $r \in A$ has the same predefined starting and ending position in the genome start(A) and end(A), respectively. Two amplicons A_1, A_2 are considered overlapped if and only if start(A_1) $\leq$ start(A_2) $<$ end(A_1) $\leq$ end(A_2). A set of amplicons $A = \{A_1,\ldots, A_m\}$ is said to be overlapping if and only if A_i and A_{i+1} overlap for i=1…m-1. Given an overlapping set A, we define a partial order $<$ on the set of reads $R = A_1 \cup \ldots \cup A_m$ as follows: $r < r'$ if and only if $r \in A_i$, $r' \in A_{i+1}$ and r and r' are consistent over their overlap of length $l_{i,i+1}$= end(A_i)- start(A_{i+1})+1, i.e., the suffix of length $l_{i,i+1}$ of r coincides with the prefix of length $l_{i,i+1}$ of r'.

Given an overlapping set $A = \{A_1,\ldots, A_m\}$, we construct an (m+2)-staged directed vertex-weighted read-graph as follows: G=($V(G)$=$V_1 \cup \ldots \cup V_m \cup \{s, t\}$, $E(G)$, c), where each $v \in V_i$, $1 \leq i \leq m$ corresponds to a distinct read $r_v \in A_i$. An edge $(u,v) \in E(G)$ if and only if either $r_u < r_v$ or $u = s$, $v \in A_1$ or $u \in A_m$, $v = t$. We also define

the function c: $V_1 \cup \ldots \cup V_m \rightarrow [0, 1]$, where $c(v)$ denotes the frequency of the read represented by $v \in V_i$ in amplicon A_i. It is evident that every full-size quasispecies that has a sequenced read from each amplicon A_i corresponds to an (s,t)-path in the graph G.

A bipartite clique of G is defined as a set of vertices $C \subseteq V(G)$ such that $C \subseteq V_i \cup V_{i+1}$ for some i and every vertex from the set $C \cap V_i$ is adjacent to every vertex from the set $C \cap V_{i+1}$.

Lemma 1. *Consistent overlaps in amplicons A_i, A_{i+1} correspond to disjoint bipartite cliques in G.*

Proof. Suppose the contrary; then there exist vertices v, $v' \in A_i$ and u, $u' \in A_{i+1}$, such that $r_v < r_u$, $r_v < r_{u'}$, $r_{v'} < r_u$, but it is not true that $r_{v'} < r_{u'}$. Since $r_{v'}$ and r_u are comparable but $r_{v'}$ and $r_{u'}$ are not, the prefixes of length $l_{i,i+1}$ of r_u and $r_{u'}$ must not be consistent. This implies a contradiction with $r_v < r_u$ and $r_v < r_{u'}$.

Using this simple finding, we transform the read graph G into a new ``forked" edge-weighted directed read-graph $H = (V(H), E(H), d)$ as follows. Consider each $p \times q$-bipartite clique $C = K_{p,q}$ of G not containing vertices s,t. $C \subseteq A_i \cup A_{i+1}$ for some $i \in \{1,\ldots,m\text{-}1\}$. Add a new "fork" vertex v_{fork}, delete all edges of the bipartite clique C and add edges from the sets $\{(\underline{u}, v_{fork}) : u \in C \cap A_i\}$ and $\{(v_{fork}, v) : v \in C \cap A_{i+1}\}$. Define a new edge weight function d : $E(H) \rightarrow N$ as follows: $d(uv_{fork}) = c(u)$, $d(v_{fork}v) = c(v)$, $d(su) = d(vt) = 0$. Figure 1 illustrates this transformation.

Figure 1 - Transformation of read graph into forked read graph

As for *G,* every full-size quasispecies corresponds to (*s,t*)-path in the forked read graph *H*. However, *H* is (2*m*+1)-staged directed graph with much fewer edges than *G*: for every bipartite clique $K_{p,q}$ pq edges in *G* are replaced by only $p+q$ edges in *H*. Since in network flow problems variables usually are associated with edges, this reduction is highly useful for the construction of the fast network flows-based algorithm for the quasispecies spectrum reconstruction problem.

The quasispecies reconstruction problem may be restated as the following covering problem:

Problem 1. Given a forked read graph *H*, cover *H* with a set of unique (*s,t*)-paths P_i with frequencies g_i such that the total frequency of paths is minimal and for every directed edge $(u,v) \in E(H)$ the sum of frequencies of paths containing (*u*,*v*) is at least *d(uv)*.

We next reformulate Problem 1 as a MCF problem. Suppose that *k* is an upper bound for the number of quasispecies (k is the parameter of the algorithm analogous to the parameters of clustering algorithms such as k-means). Then an exact solution of Problem 1 could be obtained using the following Mixed Integer Linear Programming formulation:

$$minimize \sum_{\substack{i=1,\dots k \\ (s,u)\in E(H)}} g_{su}^{i} \quad (1)$$

s.t.

$$\sum_{i=1}^{k} g_{uv}^{i} \geq d(uv), \quad (u,v) \in E(H) \quad (2)$$

$$\sum_{(u,v)\in E(H)} g_{uv}^{i} = \sum_{(v,w)\in E(H)} g_{vw}^{i}, \quad v \in V(H)\backslash\{s,t\}, \qquad i = 1,\dots,k \quad (3)$$

$$\sum_{(v,u)\in E(H)} f_{vu}^{i} \leq 1, \quad v \in V(H), \quad i = 1,\dots,k \quad (4)$$

$$f_{uv}^{i} \geq g_{uv}^{i}, \quad (u,v) \in E(H), \quad i = 1, \ldots, k \quad (5)$$

$$f_{uv}^{i} \in \{0,1\}, \qquad (u,v) \in E(H), \qquad i = 1, \ldots, k \quad (6)$$

$$g_{uv}^{i} \in [0,1], \quad (u,v) \in E(H), \quad i = 1, \ldots, k \quad (7)$$

The variables g_{uv}^{i} represent the values of the flow *i* on the edge (*u,v*). With each flow g^i we associate a binary vectors f^i such that for every $(u,v) \in E(H)$

$$\text{if } g_{uv}^{i} > 0 \text{ then } f_{uv}^{i} = 1 \qquad (8)$$

This condition is guaranteed by the constraints (5). Constraints (2) and (3) are covering and flow conservation constraint, respectively. Constraints (4) guarantee that flows g^i are unsplittable for every i=1,…,k, i.e. the edges carrying each flow forms a simple (*s,t*)- path P_i in the forked read-graph *H*. In particular, the constraint implies that for every i=1,…,k the values g_{uv}^{i} are equal for all edges of P_i. Therefore g_{uv}^{i} can be interpreted as values proportional to frequencies of quasispecies *i*.

The frequency of *i*-th quasispecies is calculated as the normalized size of the *i*-th flow by the formula

$$\frac{\sum_{su \in E(H)} g_{su}^{i}}{\sum_{i=1}^{k} \sum_{su \in E} g_{su}^{i}}. \qquad (9)$$

ShotMCF algorithm

The input is a set of distinct reads *R* with counts (c_v : $v \in R$) and a set of candidate sequences $Q=\{q_1,...,q_k\}$ generated using the max bandwidth method of ViSpA . We construct the directed read graph $G = (V, E)$ as follows:

1) for each read $r_v \in R$ aligned with the reference sequence add a vertex $v \in V$; the consensus of candidate sequences can be used as a reference;
2) the directed edge (u, v) belongs to *E* if and only if some suffix of r_u overlaps with a prefix of r_v and the two reads agree inside the overlap;

3) for each candidate sequence $q_i \in Q$ add a source s_i and a sink t_i. Add edges (s_i,v) and (v,t_i) for each vertex $v \in R$ such that r_v coincides with the prefix or suffix of q_i, respectively.

Let a read r_v of length l be aligned with a candidate sequence q_i and its alignment have j mismatches (replacements, insertions and deletions). Let p^i_v be the probability that read r_v was obtained from the sequence q_i. This probability can be estimated as

$$p_v^i = \left(\frac{\varepsilon}{3}\right)^j (1-\varepsilon)^{l-j}, \qquad (10)$$

where ε is the sequencing error rate, i.e. the probability of error per nucleotide. Note that the analogous formula was in the quasispecies theory for the calculation of the probability of mutation between two different quasispecies [22].

Using the read-graph constructed above, the quasispecies frequencies estimation problem can be formulated in terms of MCF as follows. Each (s_i,t_i)-path corresponds to some full-genome quasispecies, which can coincide with q_i with a probability depending on values p^i_v. By using p^i_v as coefficients in the MCF objective function, we arrive to the following formulation:

$$maximize \sum_{v \in V} \sum_{i=1}^{k} p_v^i g_v^i \qquad (11)$$

s.t.

$$\sum_{i=1}^{k} g_v^i \le c_v, \quad v \in V \qquad (12)$$

$$\sum_{uv \in E} g_{uv}^i = \sum_{vw \in E} g_{vw}^i, \quad v \in V \backslash \{s_i, t_i\}, \quad i = 1, \dots, k \qquad (13)$$

$$g_{uv}^i \ge 0, \quad uv \in E, i = 1, \dots, k \qquad (14)$$

Here g^i_{uv} are flow variables. $g^i_v = \sum_{uv \in E} g^i_{uv}$ are auxiliary variables used for the simplicity of notations, which represent total flow through vertices $v \in V$. The resulted

flow is fractional and can split, thus allowing for accounting read errors and mutations. (11)-(14) is a variant of MCF, where vertex capacity constraints are used instead of edge capacity constraints. Once the problem is solved, the frequency of each candidate quasispecies could be estimated using (9).

Results

In order to validate the devised methods, we used reads simulated from experimentally identified intra-host HCV variants or quasispecies.

The simulated reads were generated using individual 1734-nt sequences derived from the E1/E2 genomic region of intra-host HCV variants [16]. For each run of the algorithm, quasispecies populations were generated using 10 randomly selected sequences with randomly assigned frequencies. Quasispecies frequencies were generated according to uniform, geometric, and skewed distributions.

1) In the uniform distribution all sequences have approximately equal frequencies, which were calculated as normalized numbers of times each sequence was chosen in 1000 independent trials, where at each trial one of sequences was randomly chosen with an equal probability.
2) In the geometric distribution frequencies form a geometric progression. The frequencies were calculated by taking 10 first terms in geometric progressions and normalizing them.
3) In the skewed distribution one of the sequences has a high frequency, while the remaining sequences have uniformly low frequencies (generated as in 1)).

The read lengths followed a normal distribution with mean value of 320nt and variance of 10nt. The number of reads in each simulated data set varied from 5K to 300K for ShotMCF and from 5K to 100K for AmpMCF. Shotgun reads were

simulated using FlowSim [14]. We generated amplicons with the equal length of 320nt and the difference of 250nt between starting positions of consecutive amplicons. The starting position of each amplicon read was chosen among amplicons starting positions using a uniform distribution.

For each size of a data set and for each distribution type 11 independent simulated data sets were generated, averages of measures of algorithms quality were calculated and the statistical significance of algorithms comparison was estimated using a Kruskal-Wallis test [20].

Problems formulations (1)-(7) and (11)-(14) were solved using the IBM ILOG CPLEX solver 12.2 (www.ibm.com/software/integration/optimization/cplex-optimizer/) with the default parameters. ILP for AmpMCF was solved in parallel on 16x Intel(R) Xeon(R) CPU X5550 2.67GHz, 48 GB Memory with a running time limit 5 minutes per problem. LP for ShotMCF was solved in parallel on 24x Intel(R) Xeon(R) CPU E7450 2.40GHz, 128 GB Memory to optimality. The average running time for solving LP formulation for ShotMCF varied from 30.945 seconds with a standard deviation 11.332 seconds for 5K reads to 352.301 seconds with a standard deviation 56.861 seconds for 300K reads. The average running time for solving ILP formulation for AmpMCF varied from 110.219 seconds with a standard deviation 106.342 seconds for 5K reads to 126.270 seconds with a standard deviation 99.500 seconds for 100K reads.

P-values for a Kruskal-Wallis test were calculated using MATLAB (http://www.mathworks.com/products/matlab/).

ShotMCF algorithm.

The reconstructions obtained using ShotMCF and EM algorithms from ViSpA [7] were compared. It was shown in [7] that ViSpA with EM outperforms state-of-the-art algorithm SHORAH proposed in [6]. Since EM and ShotMCF use the same method for candidate quasispecies generation, both algorithms were evaluated for the accuracy of finding the distribution of quasispecies frequencies. Following [7] and [11], we used two measures of accuracy: Root Mean Square Error (RMSE) and Kullback-Leibler divergence (KLD) [10] between the estimated distribution and the true distribution. KLD is a quasi-metric that measures the distance between two probability distributions $P=(p_1,\ldots,p_n)$ and $W=(w_1,\ldots,w_n)$ by the following formula:

$$KLD(P,W) = \sum_{i=1}^{n} \ln\left(\frac{p_i}{w_i}\right) p_i$$

The following figures illustrates the comparison of ShotMCF and EM algorithms

Figure 2 – Comparison of ShotMCF and EM – RMSE

Figure 3 – Comparison of ShotMCF and EM – KLD

The difference in performance between two algorithms is statistically significant for all distributions and sizes of data. The p-values of a Kruskal-Wallis test are summarized in Table 1.

Table 1 - Statistical significance of the comparison of ShotMCF and EM

ShotMCF statistically significantly outperforms EM on large data sets with geometric and skewed distributions, while the quality of EM is higher on small data sets. The quality of quasispecies reconstruction by EM, as implemented in ViSpA [7], declined with the increase in the dataset size for large numbers of reads, and was not significantly affected for ShotMCF. EM produced more accurate results on data sets with up to 300K reads generated using the uniform distribution. However, the trend of decrease in quality of EM estimations suggests that ShotMCF is more accurate on the larger data sets generated using the uniform distribution.

The accuracy of frequency estimation for variants with different abundances was analysed (Fig. 8)

Figure 8 – Dependence between relative error (RE) in frequency estimation and an abundance of a variant – ShotMCF

Here, all analysed sequences were partitioned into 5 groups according to their frequencies f: $f \leq 0.025$, $0.025 < f \leq 0.05$, $0.05 < f \leq 0.1$, $0.1 < f \leq 0.2$ and $f > 0.2$. x-axis represents the groups and y-axis represents the average relative error of ShotMCF for each group. Frequencies of high-abundance variants were estimated more accurately. The accuracy of frequencies estimation increases monotonically with the abundance and stabilizes approximately at the abundance 0.1. The quality of frequency estimation increases, in general, with the number of reads in data set for all groups.

AmpMCF algorithm

The reconstructions obtained using AmpMCF (k=12) and the Maximum Bandwidth (MB) algorithm proposed in [8] were compared. Maximum bandwidth is based on the packing formulation of the quasispecies spectrum reconstruction problem, and was shown to outperform the algorithm for quasispecies spectrum reconstruction from amplicon reads proposed in [9]. The following measures of quality of a solution were used:

1) RMSE

2) Jensen-Shannon divergence (JSD). It replaces KLD used for ShotMCF testing, since for AmpMCF and MB sizes of the reconstructed quasispecies populations may differ from the size of the correct population. JSD differs from KLD divergence due to the addition of a midpoint and is defined as follows:

$$JSD(P,W) = \frac{1}{2}KLD(P,M) + \frac{1}{2}KLD(W,M),$$

where P and S are probability distributions and $M = \frac{1}{2}(P + W)$.

3) Sensitivity S, which is defined as

$$S = \frac{|TruePositives|}{|TruePositives| + |FalseNegatives|}$$

4) Positive predicted value (PPV), which is defined as

$$= \frac{|TruePositives|}{|TruePositives| + |FalsePositives|}$$

Here, if CandQ is the set of quasispecies found by the algorithm and SimQ is the set of simulated quasispecies, then TruePositives = CandQ∩SimQ, FalseNegatives = SimQ\CandQ and FalsePositives = CandQ\SimQ.

RMSE and JSD measure the quality of quasispecies frequencies estimation, and Sensitivity and PPV measure the quality of assembled quasispecies. Sensitivity is a measure of the positive identifications, which is defined as the percentage of

correctly assembled quasispecies out of the population. PPV is a measure of the negative identification, which is defined as the percentage of correctly identified quasispecies over all assembled quasispecies.

The following figures illustrates the comparison of AmpMCF and Maximum Bandwidth algorithms

Figure 4 – Comparison of AmpMCF and Maximum Bandwidth – RMSE

Figure 5 – Comparison of AmpMCF and Maximum Bandwidth – JSD

Figure 6 – Comparison of AmpMCF and Maximum Bandwidth – Sensitivity

Figure 7 – Comparison of AmpMCF and Maximum Bandwidth – PPV

The p-values of a Kruskal-Wallis test are summarized in Table 2.

Table 2 - Statistical significance of the comparison of AmpMCF and Maximum Bandwidth

According to RMSE, AmpMCF statistically significantly outperforms Maximum Bandwidth for all sizes of data sets with the geometric distributions, and for large data sets with the uniform distribution. Although AmpMCF exceeded in accuracy Maximum Bandwidth on the 5K and 20K datasets with the uniform distribution, the difference in performance was statistically insignificant, with p-value being slightly greater than the statistical significance threshold of 5%. For the skewed distribution

the results were comparable without statistically significant advantage of one algorithm over the other.

According to JSD and PPV, ShotMCF statistically significantly outperforms Maximum Bandwidth on the 100K data sets with the geometric distribution, while Maximum Bandwidth had the lower JSD values on the 20K and 100K data sets with the skewed distribution. For all other measures, sizes and distributions the results were comparable with no statistically significant advantage of one algorithm over the other. The p-value for S could not be calculated for the 5K data sets with the skewed distribution, since both algorithms were equally sensitive on all test examples.

So AmpMCF outperformed Maximum Bandwidth in quasispecies frequencies estimation for populations with geometric and uniform distributions, while both algorithms showed a similar performance in quasispecies sequence reconstruction.

The low sensitivity of AmpMCF and Maximum Bandwidth on the 5K data set with the skewed distribution is associated with the erroneous reconstruction of low-abundance variants by both algorithms, with only a dominant variant being correctly identified. For larger data sets, populations with the skewed distributions were reconstructed much more successfully and variants with frequencies as low as 0.8% were detected. It should be also noted that low-frequency variants were detected with higher probability in populations with the geometric distribution (Fig. 10). It suggests that the recoverability of low-frequency variants depends on the structure of a population and that the coverage provided by data sets of 5K reads is insufficient for low-frequency variants detection, if the population contains a dominant high-frequency variant.

Figure 10 – Probabilities of detection of low-frequency variants (< 0.025) for the geometric and skewed distributions - AmpMCF

In general, abundances of variants greatly affect their recoverability, with high-frequency variants being easier to detect (Fig. 11). As above, all analysed sequences in Fig. 11 were partitioned into 5 groups according to their frequencies f: $f \leq 0.025$, $0.025 < f \leq 0.05$, $0.05 < f \leq 0.1$, $0.1 < f \leq 0.2$ and $f > 0.2$. x-axis represents the groups and y-axis the probability of variant recovery in each group. The probabilities of detection of variants within each group increase with the number of reads in a data set. While the probability of reconstruction of a variant with frequency less than 2.5% from the 5K data set was only 0.0092, all variants with frequencies greater than 20% were reconstructed from 20K and 100K data sets.

Figure 11 – Probabilities of detection of quasispecies depending on their frequencies - AmpMCF

The accuracy of frequency estimation for detected variants with different abundances is illustrated by Fig. 9

Figure 9 – Dependence between relative error (RE) in frequency estimation and an abundance of a variant – AmpMCF

As for ShotMCF, the accuracy of frequency estimation increases with the abundance and stabilizes approximately at the abundance 0.1. In general, the accuracy of frequency estimation increases with the number of reads in a data set for all groups

except for the group of low-frequency variants. The small value of RE for low-frequency variants from the 5K data sets can be explained with a low detection rate of such variants, which renders their RE undefined.

Discussion

Two different network-flows based formulations applicable to quasispecies frequency reconstruction problem were developed. The first quasispecies spectrum reconstruction method based on network flows (NF) was proposed in [15]. However, usage of NF in that method does not allow the direct reconstruction of quasispecies sequences and their frequencies. Rather, it selects pairs of overlapping reads that belong to the same sequence variant. For the direct quasispecies spectrum reconstruction the second stage of the algorithm was proposed, which involves finding edge-disjoint paths in the network modified according to the results of the NF stage. The network modification substantially increases the number of edges; therefore, since edge-disjoint paths problem is NP-complete [24], the method is computationally extensive.

AmpMCF extends the concept developed in [15]. It replaces NF with MCF, which allows for joining both stages of algorithm from [15] in a single MCF formulation and for solving it using a single algorithm. Such approach is more effective and allows for increasing quality of the solution. Moreover, instead of increasing the size of the network, AmpMCF allows to decrease it, thus making the problem much more computationally tractable.

ShotMCF extends the ViSpA algorithm described in [7]. The method proposed in [7] consists of two stages: generation of candidate quasispecies sequences from shotgun NGS reads using Maximum Bandwidth paths in the read graph and

estimation of their frequencies using the Expectation Maximization (EM) algorithm [23]. ShotMCF models and solves the quasispecies frequency estimation problem using MCF instead of EM. Unlike AmpMCF and the algorithm from [15], it is a packing algorithm that invokes the vertex rather than edge capacity constraints and does not require integer variables. This new method in combination with the candidate sequences generation algorithm from [7] presents a novel framework for the reliable reconstruction of quasispecies spectrum.

The formulation for AmpMCF could not be applied to shotgun data since it assumes that each full-length sequence corresponds to a unique (s,t)-path in the read graph. However, it is not true for the shotgun data since certain sequences can be assembled from reads through different paths. This observation taken together with consideration of the structure of the read graph described by Lemma 1 indicates that the formulation is more suitable for amplicons. The analogue of AmpMCF for a shotgun data is the NF-based algorithm from [7]. However, as aforementioned, it is computationally extensive and known to be outperformed by ViSpA.

Although the formulation of ShotMCF is applicable to amplicons, AmpMCF is more suitable for this task since ShotMCF handles only the second stage of quasispecies spectrum reconstruction problem, with the first stage being the candidate sequence generation adopted from ViSpA; while AmpMCF incorporates the whole problem into a single formulation.

The structure of the read graph explains a better match of the amplicon data to the covering rather than to packing formulation implemented by Maximum Bandwidth. According to Lemma 1, consistent overlaps between consecutive amplicons form bipartite cliques in a read graph. Edges within each bipartite clique are equal in respect to choosing (s,t)-paths in a read graph. It leads to a large number

of peer alternatives for quasispecies assembling, indicating the need in search for the most parsimonious solution. The NF-based formulation with parsimony as an objective function and without predefined flow sizes requires covering constraints, and, therefore, leads to the covering formulation.

The advantage of ShotMCF method over EM-based method of ViSpA originates from enforcing uniformity of quasispecies coverage and using more accurate formula for the probability of emission of a given read from a given candidate sequence. The major advantage of the EM algorithm over the current version of ShotMCF is a greater speed and reduced requirements for computational resources such as computer memory and number of parallel processors. The reason is that MCF is implemented directly using linear programming formulation. It is expected that application of faster methods; e.g., based on lagrangian relaxations or Bender decomposition, should dramatically increase performance of ShotMCF.

It should be noted that MCF formulations assume absence of gaps in coverage. Although such gaps interrupt the assembly of the entire sequence, the genomic regions covered with reads can be identified using a reference sequence and quasispecies can be estimated with MCF-based algorithms for each region independently

Conclusions

Two novel methods were developed for the reconstruction of the structure of viral population from the NGS shotgun and amplicon reads. Both methods are based on MCF and found suitable for the reliable assembly of viral quasispecies and estimation of their frequencies.

Competing interests

Authors declare that they have no competing interests.

Authors' contributions

PS developed the algorithms and wrote the manuscript. NM developed, implemented and tested AmpMCF algorithm. AA developed, implemented and tested ShotMCF algorithm. BT prepared the testing data. IM contributed to designing the algorithms and writing the manuscript. YK contributed to designing the algorithms and writing the manuscript. AZ developed the algorithms, wrote the manuscript and supervised the project. All authors read and approved the final manuscript.

Acknowledgements

This work has been partially supported by Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture and NSF awards IIS-0916401 and IIS-0916948.

Authors thank referees for valuable comments which helped to significantly improve the paper.

References

1. Domingo E: **Biological significance of viral quasispecies.** *Viral Hepatitis Rev. 2,* 1996, 247-261.
2. Zagordi O, Klein R, Däumer M, Beerenwinkel N: **Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies**. *Nucleic Acids Research* 2010, **38**(21):7400-7409.
3. Duarte EA, Novella IS, Weaver SC, Domingo E, Wain-Hobson S, Clarke DK, Moya A, Elena SF, de la Torre JC, Holland JJ.: **RNA Virus Quasispecies: Significance for Viral Disease and Epidemiology. Infectious Agents and Disease** 3(4) (1994) 201–214.
4. Skums Pavel, Dimitrova Zoya, Campo David S., Vaughan Gilberto, Rossi Livia, Forbi Joseph C, Yokosawa Jonny, Zelikovsky Alex, Khudyakov Yury: **Efficient error correction for next-generation sequencing of viral amplicons**. *BMC Bioinformatics* 2012, **13**(Suppl 10):S6
5. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N: **Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction.** . *Journal of Computational Biology* 2009, **17**(417-428).
6. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N: **ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data**. *BMC Bioinformatics* 2011, **12**(119).
7. Astrovskaya I., Tork, B., Mangul, S., Westbrooks, K., Mandoiu, I., Balfe, P., Zelikovsky A: **Inferring Viral Quasispecies Spectra from 454 Pyrosequencing Reads**. *BMC Bioinformatics* , **12** (Suppl 6):S1, 2011
8. N. Mancuso and B. Tork and P. Skums and I. Mandoiu and A. Zelikovsky: **Viral Quasispecies Reconstruction from Amplicon 454 Pyrosequencing Reads**. *In Silico Biology* **11** (2011/2012) 1–13
9. Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, SolmoneMC, Capobianchi MR, Ulivi G: **Combinatorial Analysis and Algorithms for Quasispecies Reconstruction using Next-Generation Sequencing**. *BMC Bioinformatics* 12:5 (2011)
10. Kullback S., Leibler R.A. **On information and sufficiency** *The Annals of Mathematical Statistics*. 1951. V.22. N. 1. P. 79-86.
11. 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.

12. Quince,C. et al. **Removing noise from pyrosequenced amplicons.** *BMC Bioinformatics* 2011, 12:38.
13. Skums P, Campo D, Dimitrova Z, Vaughan G, Lau D, Khudyakov Y: **Numerical detection, measuring and analysis of differential interferon resistance for individual HCV intra-host variants and its influence on the therapy response**. *In Silico Biology* 11 (2011/2012) 1–7.
14. S. Balser, K. Malde, A. Lanzen, A. Sharma, and I. Jonassen. **Characteristics of 454 pyrosequencing data-enabling realistic simulation with FlowSim**. *Bioinformatics* 2010, 26:i420-5.
15. Westbrooks K, Astrovskaya I, Campo D, Khudyakov Y, Berman P, Zelikovsky A: **HCV Quasispecies Assembly using Network Flows**. In: Proc. International Symposium Bioinformatics Research and Applications. (2008) 159–170
16. 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.
17. Luciani F, Bull RA, Lloyd AR. **Next generation deep sequencing and vaccine design: today and tomorrow.** *Trends Biotechnol*. 2012 Sep;30(9):443-52
18. Beerenwinkel N, Zagordi O. **Ultra-deep sequencing for the analysis of viral populations.** *Curr Opin Virol*. 2011 Nov;**1**(5):413-8.
19. Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. **Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data**. *Front Microbiol*. 2012;3:329
20. Kruskal W, Wallis W. **Use of ranks in one-criterion variance analysis**, *Journal of the American Statistical Association* 1952, **47** (260): 583–621
21. R. K. Ahuja, T. L. Magnanti, J. B. Orlin: *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 1993.
22. M. Nowak: *Evolutionary dynamics*, Belknap Press of Harvard University Press, 2006.
23. Dempster, A.P.; Laird, N.M.; Rubin, D.B. **Maximum Likelihood from Incomplete Data via the EM Algorithm**. *Journal of the Royal Statistical Society*. Series B (Methodological) 1977; 39 (1): 1–38.
24. Garey, M. R. and Johnson, D. S: *Computers and Intractability. A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco, 1979.

Tables

Table 1 - Statistical significance of the comparison of ShotMCF and EM

				Geometric distribution			
# of reads	5000	20000	100000	150000	200000	250000	300000
p-value RMSE	0.000071	0.000071	0.000071	0.002263	0.000122	0.000160	0.000093
p-value KLD	0.000071	0.000913	0.000071	0.038598	0.006428	0.016540	0.000071
				Uniform distribution			
# of reads	5000	20000	100000	150000	200000	250000	300000
p-value RMSE	0.000071	0.000071	0.000071	0.000443	0.000566	0.001449	0.005258
p-value KLD	0.000071	0.000071	0.000122	0.000345	0.000566	0.001449	0.005258
				Skewed distribution			
# of reads	5000	20000	100000	150000	200000	250000	300000
p-value RMSE	0.000071	0.000071	0.027823	0.000071	0.000720	0.000093	0.000071
p-value KLD	0.000071	0.000071	0.027823	0.000071	0.001152	0.001152	0.000071

Table 2 - Statistical significance of the comparison of AmpMCF and EM

	Geometric distribution		
# of reads	5000	20000	100000
p-value RMSE	0.001100	0.000069	0.000070
p-value JSD	0.200130	0.742240	0.000718
p-value S	0.46294	0.11743	0.84517

p-value PPV	0.66827	0.79078	0.037853
	Uniform distribution		
# of reads	5000	20000	100000
p-value RMSE	0.122800	0.061063	0.015030
p-value JSD	0.339790	0.818120	0.742170
p-value S	0.34978	0.78918	0.89135
p-value PPV	0.13832	0.89501	0.50755
	Skewed distribution		
# of reads	5000	20000	100000
p-value RMSE	0.469220	0.717980	0.224440
p-value JSD	0.211260	0.004284	0.023486
p-value S	-	0.12341	0.39881
p-value PPV	0.20846	0.53018	0.40896

Figures

Figure 1 - Bipartite cliques in the read graph are replaced by forks.

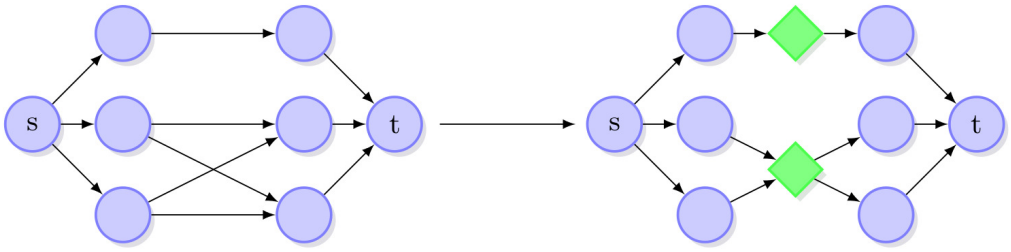
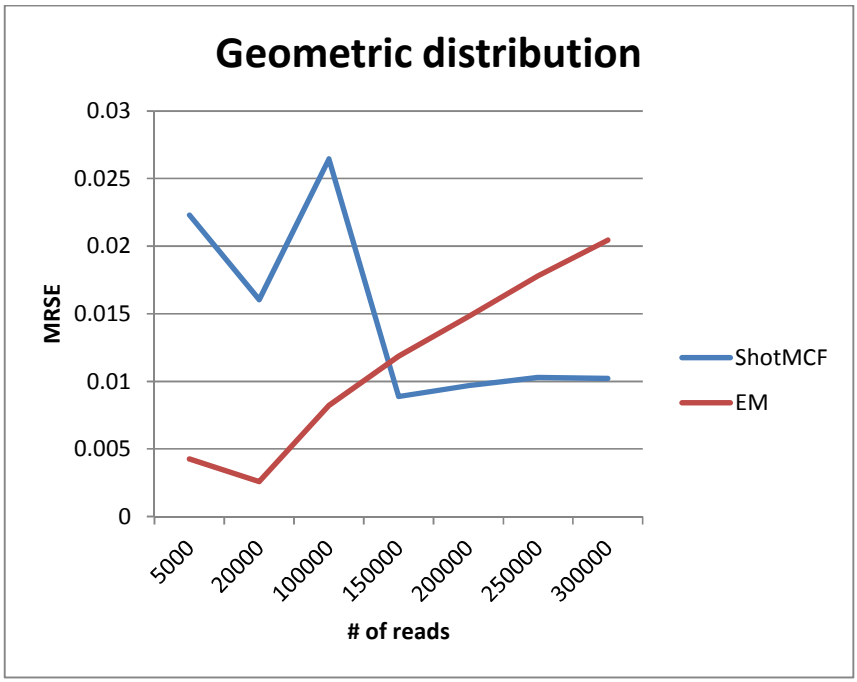
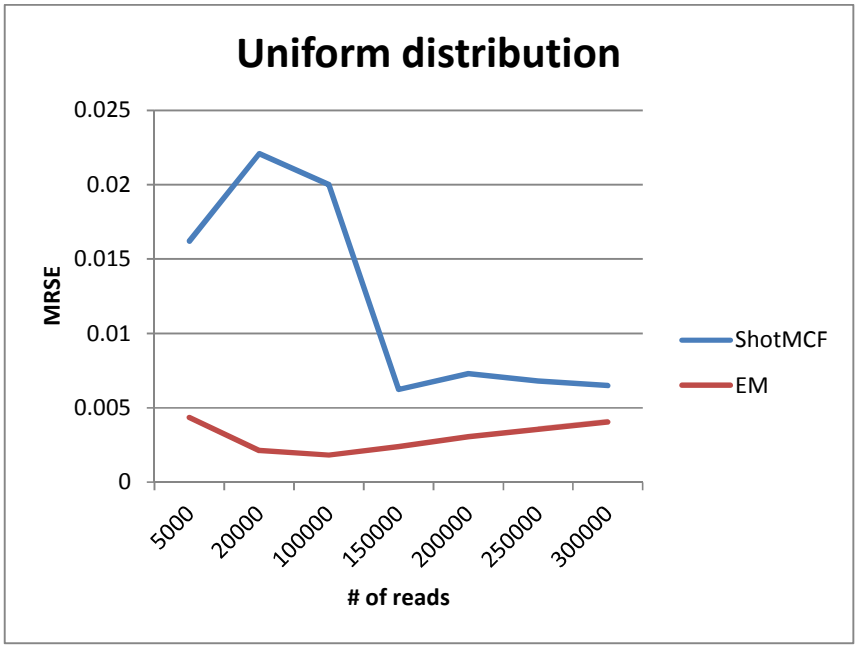


Figure 2 – Comparison of ShotMCF and EM – RMSE


Geometric distribution
0.03
0.025
0.02
0.015
0.01
0.005
0
MRSE
5000
20000
100000
150000
200000
250000
300000
# of reads
ShotMCF
EM


Uniform distribution
0.025
0.02
0.015
0.01
0.005
0
MRSE
5000
20000
100000
150000
200000
250000
300000
# of reads
ShotMCF
EM

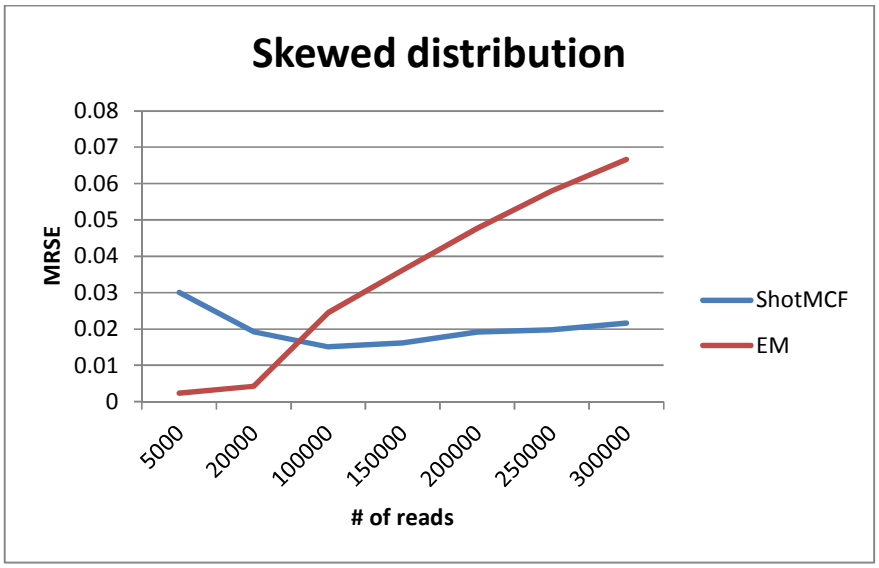
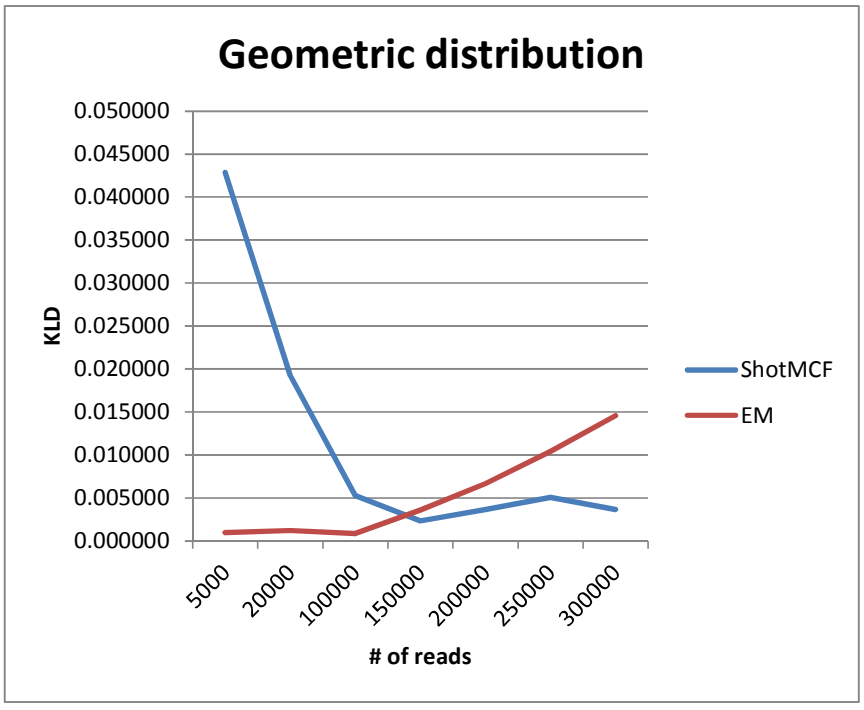


Figure 3 – Comparison of ShotMCF and EM - KLD



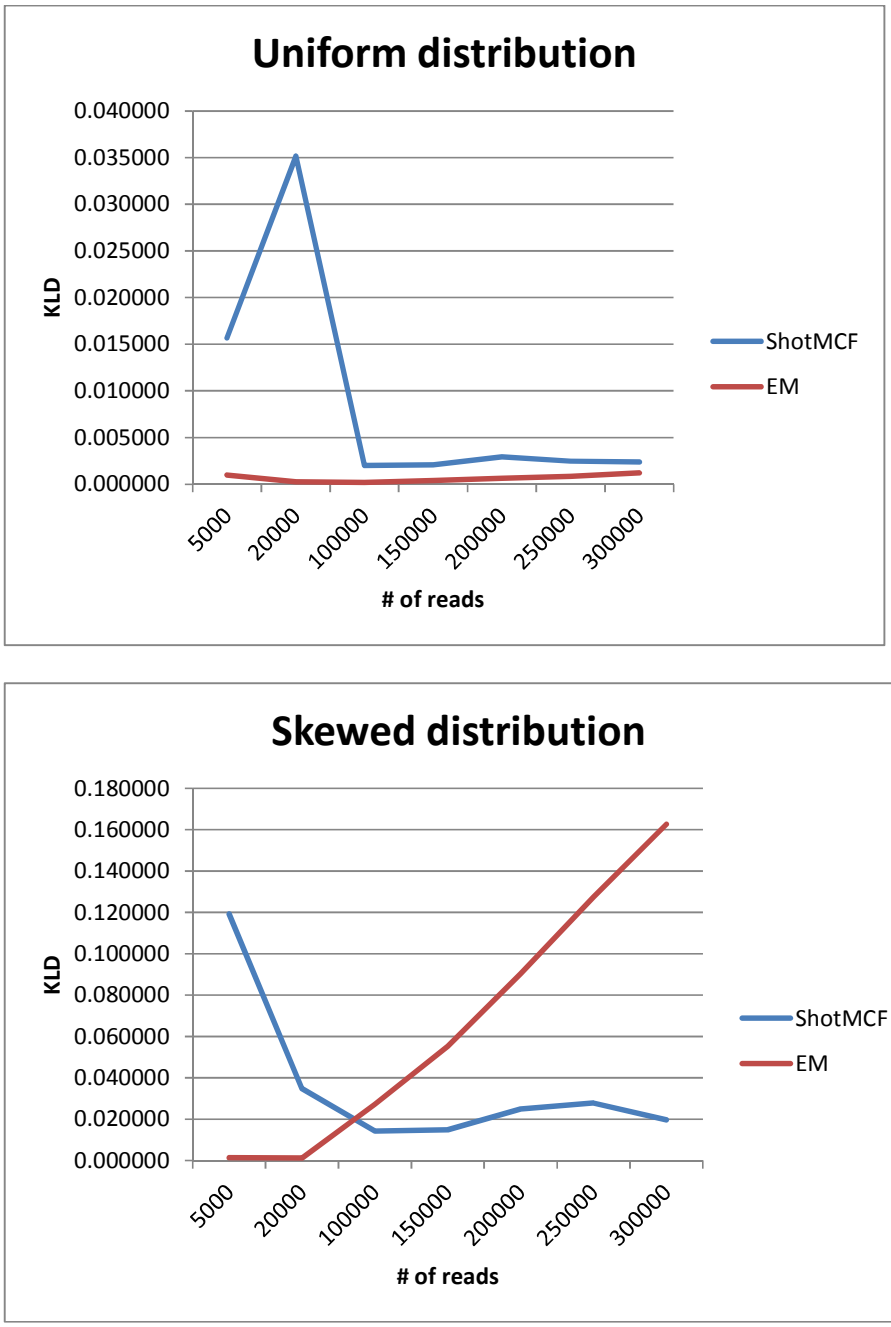
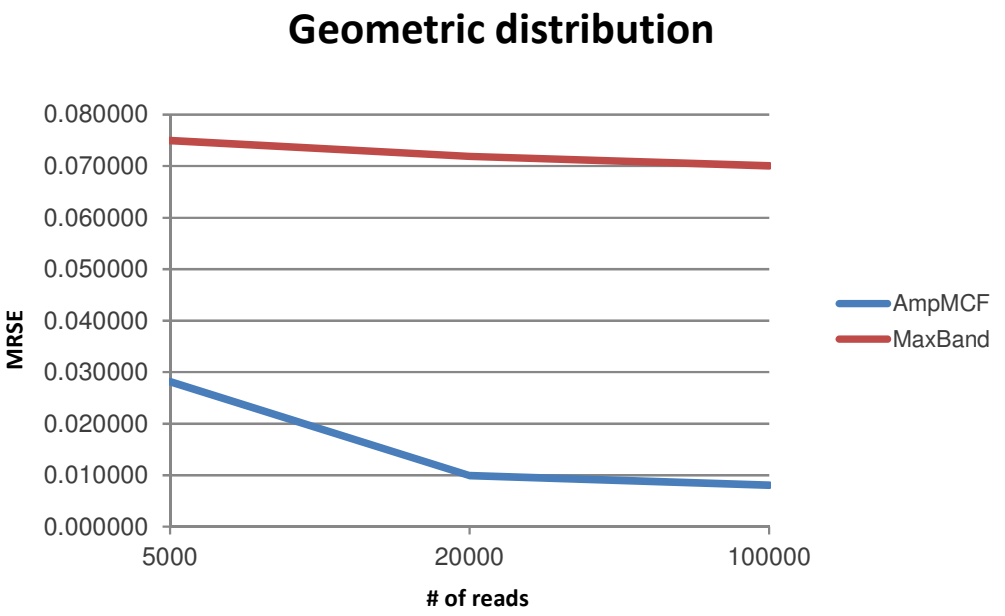


Figure 4 – Comparison of AmpMCF and Maximum Bandwidth – RMSE


Geometric distribution
0.080000
0.070000
0.060000
0.050000
0.040000
0.030000
0.020000
0.010000
0.000000
MRSE
5000
20000
100000
# of reads
AmpMCF
MaxBand


Uniform distribution
0.030000
0.025000
0.020000
0.015000
0.010000
0.005000
0.000000
MRSE
5000
20000
100000
# of reads
AmpMCF
MaxBand

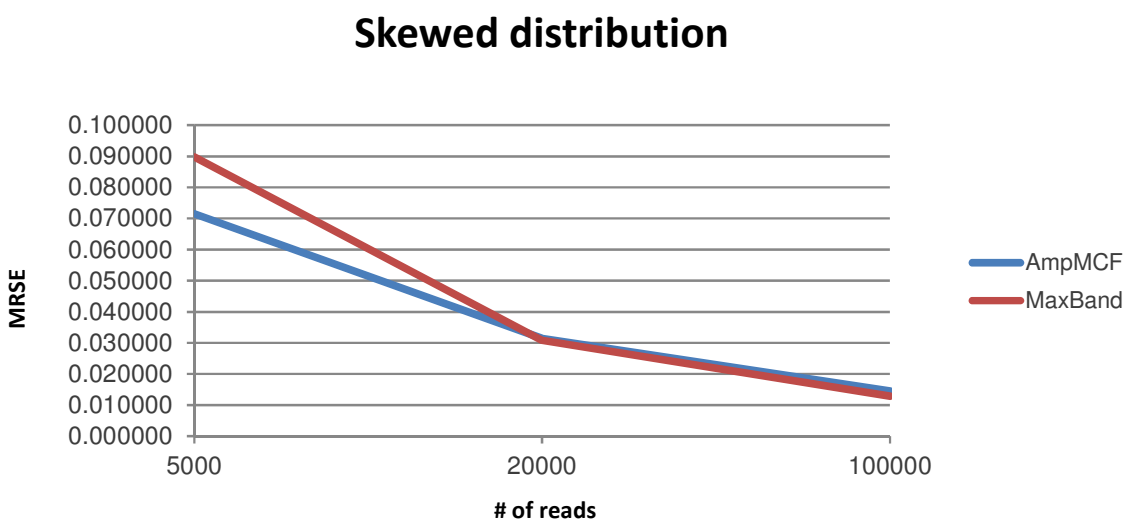
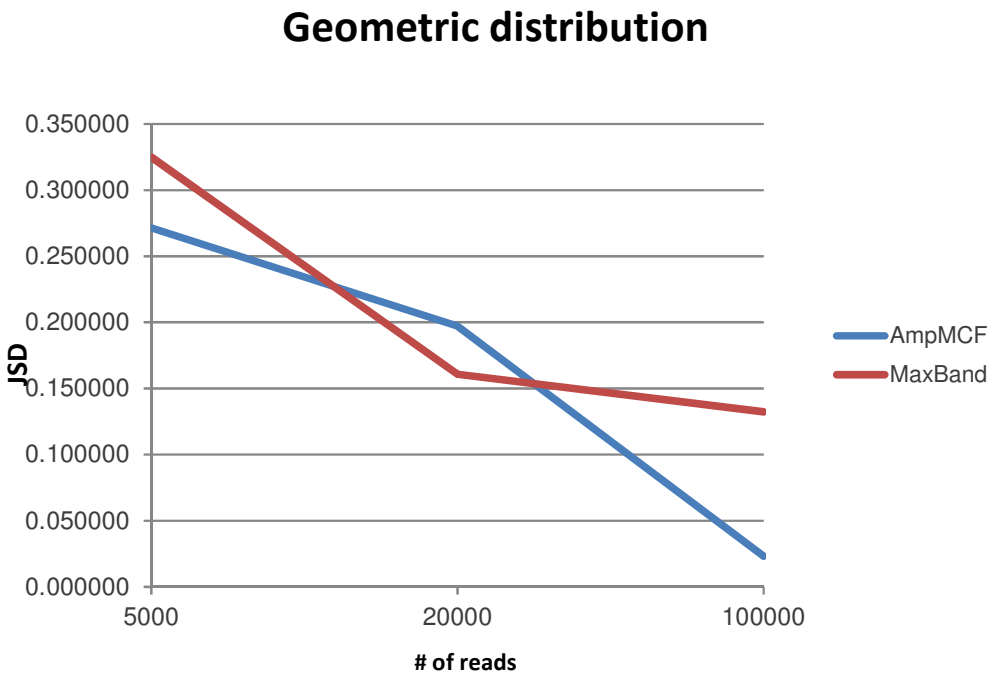


Figure 5 – Comparison of AmpMCF and Maximum Bandwidth – JSD



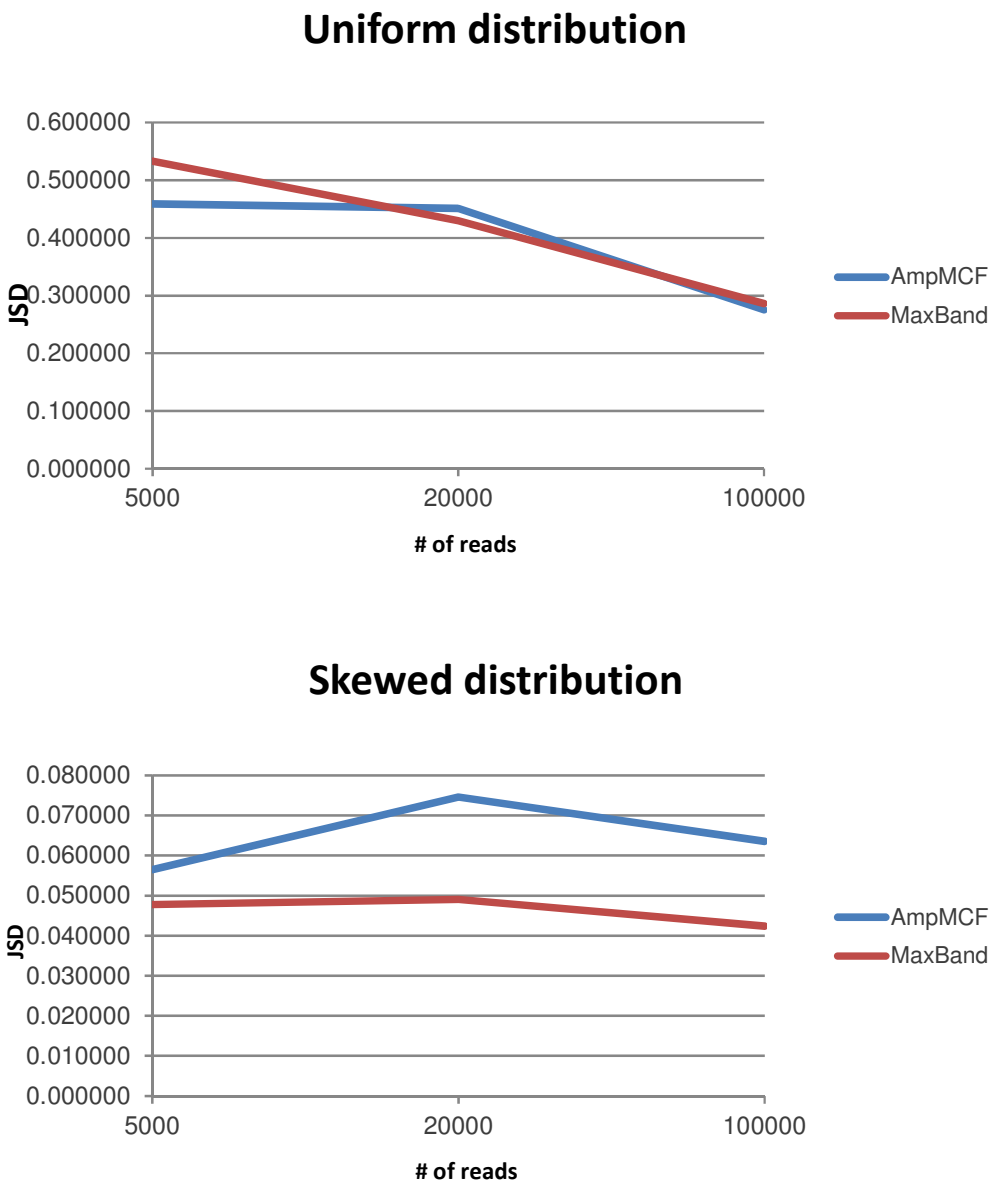
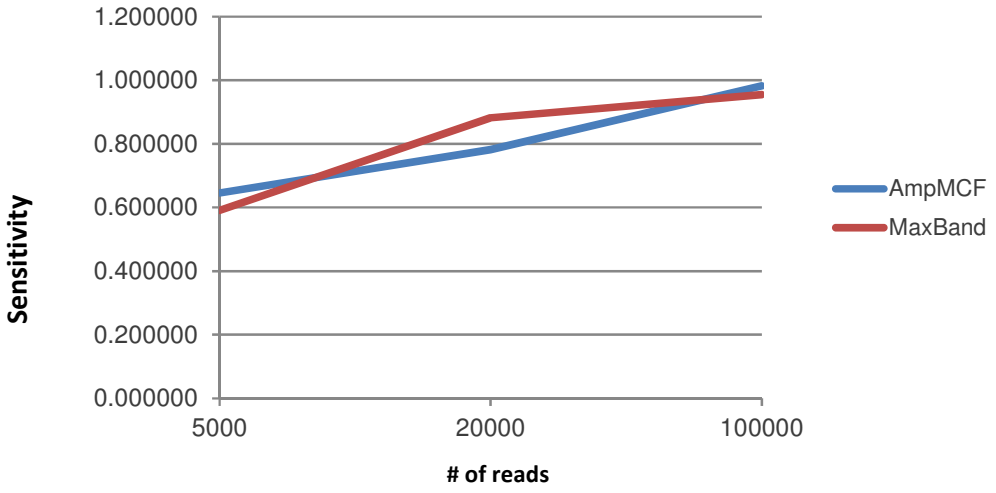
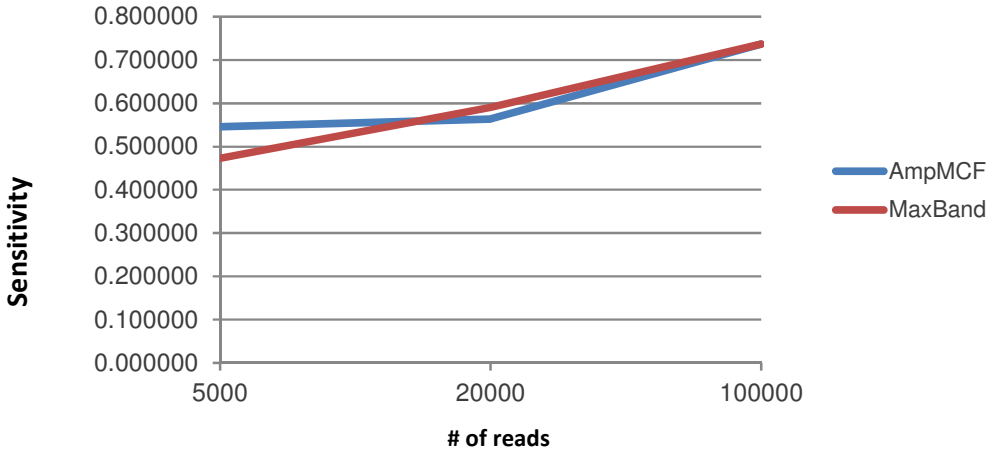


Figure 6 – Comparison of AmpMCF and Maximum Bandwidth – Sensitivity

Geometric distribution



Uniform distribution



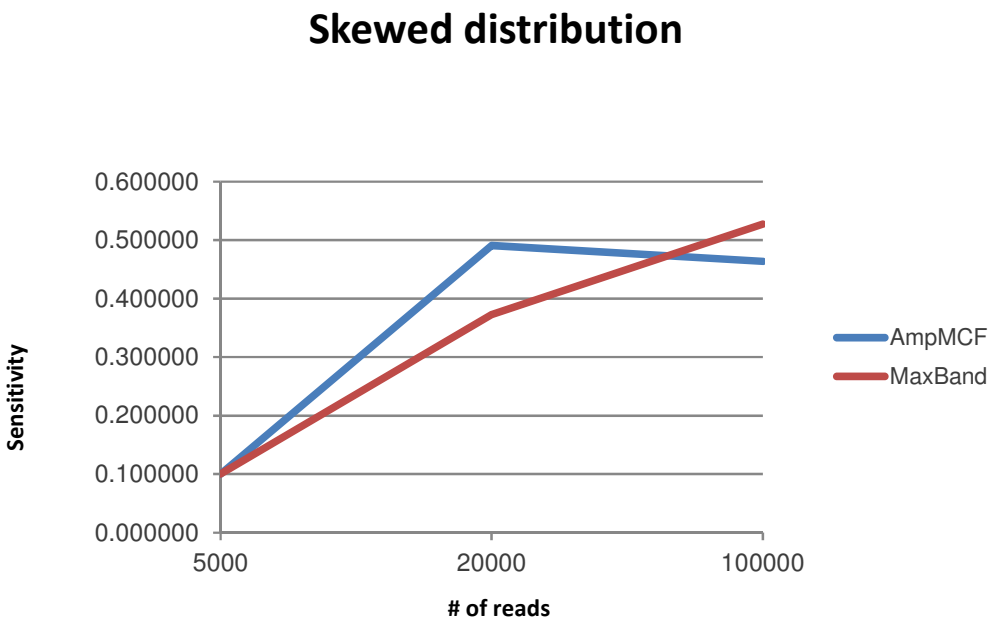
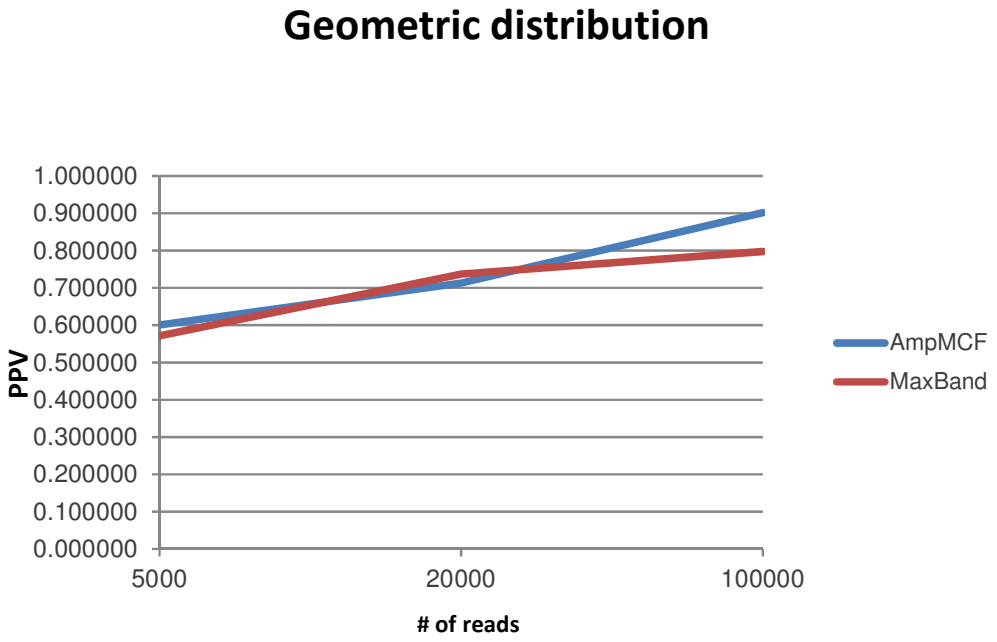


Figure 7 – Comparison of AmpMCF and Maximum Bandwidth – PPV



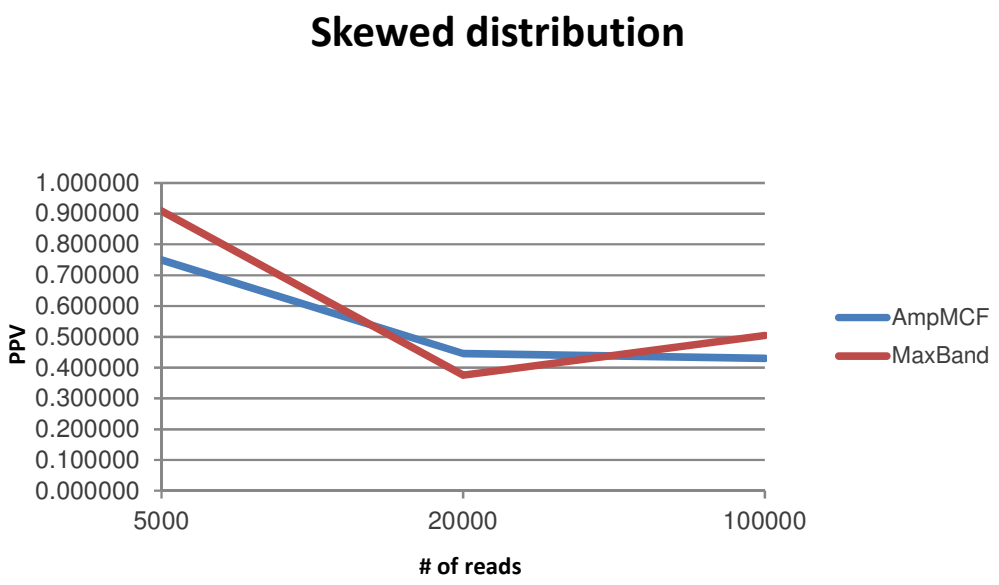


Figure 8 – Dependence between relative error in frequency estimation and an abundance of a variant - ShotMCF

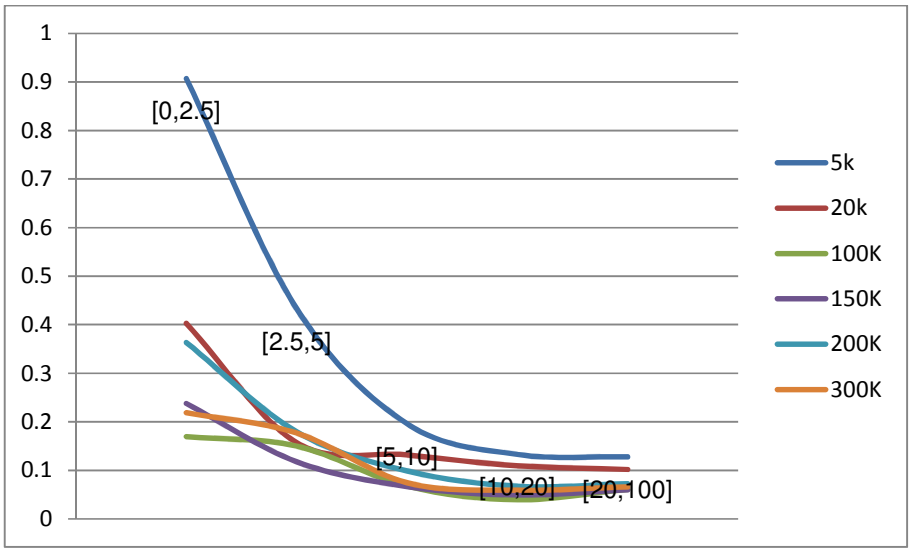


Figure 9 – Dependence between relative error in frequency estimation and an abundance of a variant – AmpMCF

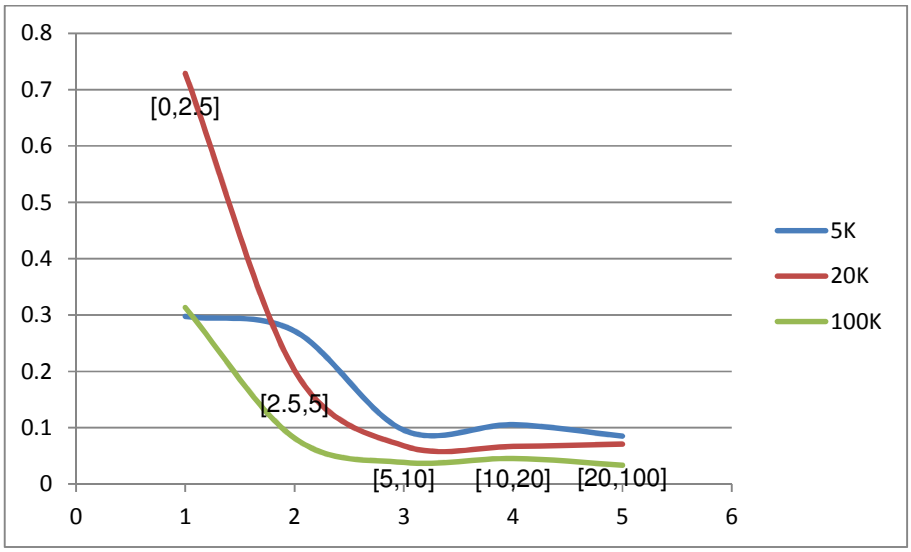


Figure 10 – Probabilities of detection of low-frequency variants (< 0.025) for the geometric and skewed distributions - AmpMCF

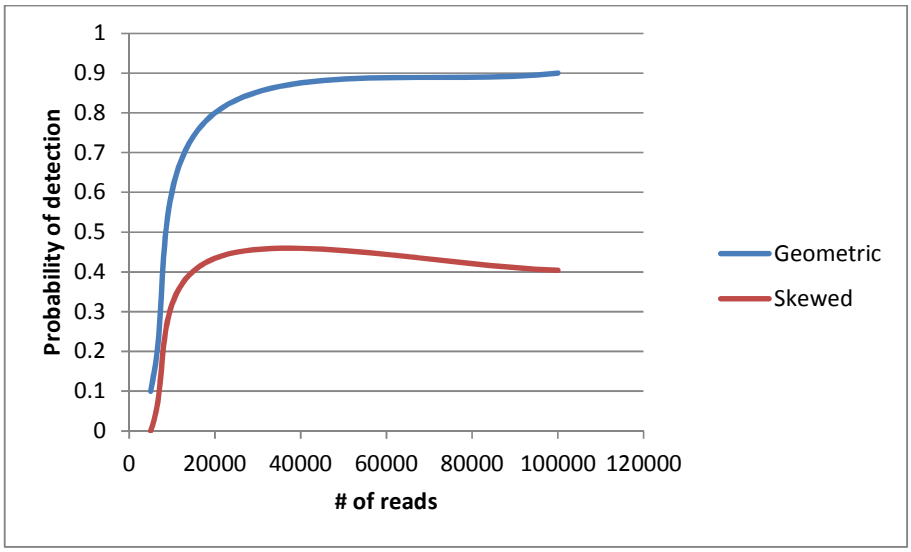


Figure 11 – Probabilities of detection of quasispecies depending on their frequencies

- AmpMCF