BIOINFORMATICS

Vol. 30 ISMB 2014, pages i329–i337
doi:10.1093/bioinformatics/btu295

Accurate viral population assembly from ultra-deep sequencing data

Serghei Mangul[1,*,†], Nicholas C. Wu[2,†], Nicholas Mancuso[3], Alex Zelikovsky[3], Ren Sun[2] and Eleazar Eskin[1,4,*]

[1]Computer Science Department, [2]Department of Molecular and Medical Pharmacology, University of California, Los Angeles, CA 90095, USA, [3]Department of Computer Science, Georgia State University, Atlanta, GA, 30303 and [4]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

ABSTRACT

Motivation: Next-generation sequencing technologies sequence viruses with ultra-deep coverage, thus promising to revolutionize our understanding of the underlying diversity of viral populations. While the sequencing coverage is high enough that even rare viral variants are sequenced, the presence of sequencing errors makes it difficult to distinguish between rare variants and sequencing errors.

Results: In this article, we present a method to overcome the limitations of sequencing technologies and assemble a diverse viral population that allows for the detection of previously undiscovered rare variants. The proposed method consists of a high-fidelity sequencing protocol and an accurate viral population assembly method, referred to as Viral Genome Assembler (VGA). The proposed protocol is able to eliminate sequencing errors by using individual barcodes attached to the sequencing fragments. Highly accurate data in combination with deep coverage allow VGA to assemble rare variants. VGA uses an expectation–maximization algorithm to estimate abundances of the assembled viral variants in the population. Results on both synthetic and real datasets show that our method is able to accurately assemble an HIV viral population and detect rare variants previously undetectable due to sequencing errors. VGA outperforms state-of-the-art methods for genome-wide viral assembly. Furthermore, our method is the first viral assembly method that scales to millions of sequencing reads.

Availability: Our tool VGA is freely available at http://genetics.cs.ucla.edu/vga/

Contact: serghei@cs.ucla.edu; eeskin@cs.ucla.edu

1 INTRODUCTION

Human immunodeficiency virus (HIV) exhibits high genomic diversity within an infected host, which affects many clinically important phenotypic traits such as escape from vaccine-induced immunity, virulence and response to antiviral therapies (Lauring and Andino, 2010). To accurately characterize an intra-host HIV population, sequencing technologies must be sensitive enough to detect and quantify rare variants (Henn *et al.*, 2012; Tsibris *et al.*, 2009). Next-generation sequencing (NGS) technologies offer deep coverage of genomic data in the form of millions of sequencing reads (Metzker, 2009). While the sequencing coverage is high enough to capture rare variants, the presence of sequencing errors makes it difficult to distinguish between rare variants and sequencing err0ors. Additionally, low viral population variability (i.e. pairs of individual viral genomes that have small genetic distance) and the presence of individual variants having low abundance complicates accessing viral diversity and assembling full-length viral variants.

The full picture of viral diversity in a population remains undiscovered due to errors produced by sequencing platforms. Current sequencing technologies use different underlying chemistry and offer trade-offs among throughput, read length and cost (Metzker, 2009). While the current sequencing platforms can potentially detect point-mutations, error rates may result in false-positive single nucleotide variant (SNV) calls or wrong genome variant sequences. Computational error correction techniques are able to partially correct the sequencing error and provide an opportunity to discover highly expressed individual viral genomes, but low abundant variants remain undiscovered. Current methods (Astrovskaya, 2011; Mancuso *et al.*, 2011; Prosperi and Salemi, 2012; Zagordi *et al.*, 2011, 2012) are not able to differentiate true biological mutations from sequencing artifacts, thus significantly limiting the possibility of a method to assemble the underlying viral population.

In this article, we propose a method to overcome these limitations by coupling a high-fidelity sequencing protocol (Kinde *et al.*, 2011) with an accurate method, referred to as **V**iral **G**enome **A**ssembler (VGA), to assemble a heterogeneous viral population. High-fidelity sequencing protocol, known as Safe-SeqS, has been applied to detect rare somatic mutations, but its application on detecting rare viral mutations has been neglected. Similar to Safe-SeqS we apply a special library preparation technique that eliminates sequencing errors during the demultiplexing step. The proposed protocol attaches individual barcode sequences during the library preparation step for every fragment, then amplifies each tagged fragment. Reads are clustered according to the original fragment based on the attached barcode. An error-correction protocol is then applied for every read group resulting in a method that corrects errors inside the group and produces a corrected consensus read. Highly accurate data in combination with deep coverage allows for accurate estimation of the underlying diversity of a viral population. Importantly, the low per-base sequencing cost of the Illumina platform makes it realistic to greatly increase coverage to detect ultra-rare variants. Our sequencing protocol introduces novel challenges for virus assembly and we develop a novel assembly approach for reconstructing and estimating the frequency of a

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

large number of closely related viral variants. Our method does not rely on the availability of a reference genome. This makes our method applicable to newly emerged viruses in which genome sequences are unknown.

The ability to discover rare viral variants makes our tool applicable for monitoring and quantifying an HIV population structure to dissect its evolutionary landscape and study genomic interaction. In particular, our approach allows for the discovery of rare mutations and variants that are of particular interest because of their potential influence on drug resistance and treatment failure (Liu *et al.*, 2011; Palmer *et al.*, 2006; Wang *et al.*, 2007).

2 METHODS

2.1 Overview

Advances in NGS and the ability to generate deep coverage data in the form of millions of reads provide exceptional resolution for studying the underlying genetic diversity of complex viral populations. However, errors produced by most sequencing protocols complicate distinguishing between true biological mutations and technical artifacts that confound detection of rare mutations and rare individual genome variants. A common approach is to use post-sequencing error correction techniques able to partially correct the sequencing errors. In contrast to clonal samples, the post-sequencing error correction methods are not well suited for mixed viral samples and may lead to filtering out true biological mutations. For this reason, current viral assembly methods are able to detect only highly abundant SNV, thus limiting the discovery of rare viral genomes.

Additional difficulty arises from the genomic architectures of viruses. Long common regions shared across viral population (known as conserved regions) introduce ambiguity in the assembly process. Conserved regions may be due low-diversity population or due to recombination with multiple cross-overs. In contrast to repeats in genome assembly, conserved regions may be phased based on relative abundances of viral variants. Low-diversity viral populations in which all pairs of individual genomes within a viral population have a small genetic distance from each other may represent additional challenges for the assembly procedure.

We apply a high-fidelity sequencing protocol to study viral population structure (Fig. 1). This protocol is able to eliminate errors from sequencing data by attaching individual barcodes during the library preparation step. After the fragments are sequenced, the barcodes identify clusters of reads that originated from the same fragment, thus facilitating error correction. Given that many reads are required to sequence each fragment, we are trading off an increase in sequence coverage for a reduction in error rate. Prior to assembly, we utilize the *de novo* consensus reconstruction tool, Vicuna (Yang *et al.*, 2012), to produce a linear consensus directly from the sequence data. This approach offers more flexibility for samples that do not have 'close' reference sequences available. Traditional assembly methods (Gnerre *et al.*, 2011; Luo *et al.*, 2012; Zerbino and Birney, 2008) aim to reconstruct a linear consensus sequence and are not well-suited for assembling a large number of highly similar but distinct viral genomes. We instead take our ideas from haplotype assembly methods (Bansal and Bafna, 2008; Yang *et al.*, 2013), which aim to reconstruct two closely related haplotypes. However, these methods are not applicable for assembly of a large (*a priori* unknown) number of individual genomes. Many existing viral assemblers estimate local population diversity and are not well suited for assembling full-length quasi-species variants spanning the entire viral genome. Available genome-wide assemblers able to reconstruct full-length quasi-species variants are originally designed for low throughput and are impractical for high throughput technologies containing millions of sequencing reads.

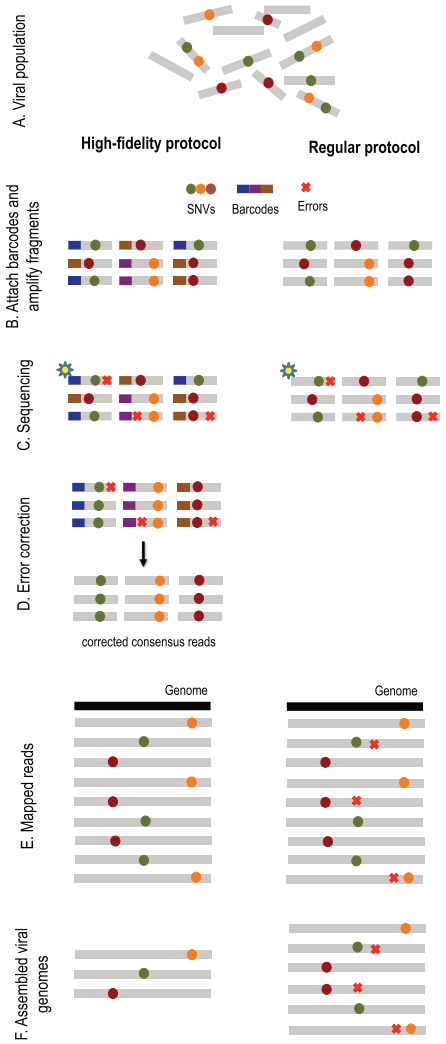

Fig. 1. Overview of high-fidelity sequencing protocol. (**A**) DNA material from a viral population is cleaved into sequence fragments using any suitable restriction enzyme. (**B**) Individual barcode sequences are attached to the fragments. Each tagged fragment is amplified by the polymerase chain reaction (PCR). (**C**) Amplified fragments are then sequenced. (**D**) Reads are grouped according to the fragment of origin based on their individual barcode sequence. An error-correction protocol is applied for every read group, correcting the sequencing errors inside the group and producing corrected consensus reads. (**E**) Error-corrected reads are mapped to the population consensus. (**F**) SNVs are detected and assembled into individual viral genomes. The ordinary protocol lacks steps (B) and (D)

We introduce a viral population assembly method (Fig. 2) working on highly accurate sequencing data able to detect rare variants and tolerate conserved regions shared across the population. Our method is coupled with post-assembly procedures able to detect and resolve ambiguity raised from long conserved regions using expression profiles (Fig. 2F). After a consensus has been reconstructed directly from the sequence data, our method detects SNVs from the aligned sequencing reads. Read overlapping is used to link individual SNVs and distinguish between genome variants in the population. The viral population is condensed in a conflict graph built from aligned sequencing data. Two reads are originated from

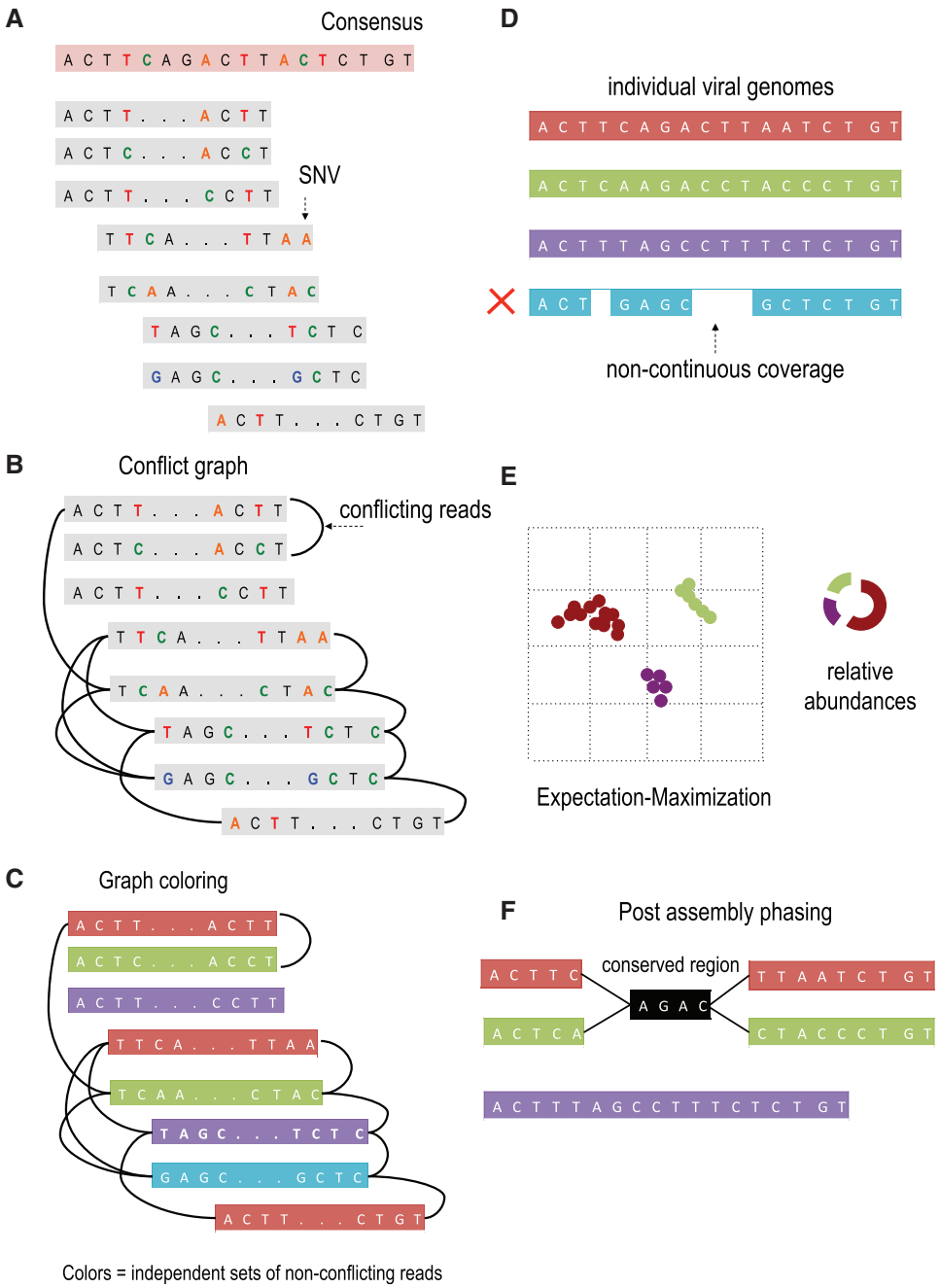

Fig. 2. Overview of VGA. (**A**) The algorithm takes as input paired-end reads that have been mapped to the population consensus. (**B**) The first step in the assembly is to determine pairs of conflicting reads that share different SNVs in the overlapping region. Pairs of conflicting reads are connected in the 'conflict graph'. Each read has a node in the graph, and an edge is placed between each pair of conflicting reads. (**C**) The graph is colored into a minimal set of colors to distinguish between genome variants in the population. Colors of the graph correspond to independent sets of non-conflicting reads that are assembled into genome variants. In this example, the conflict graph can be minimally colored with four colors (red, green, violet and turquoise), each representing individual viral genomes. (**D**) Reads of the same color are then assembled into individual viral genomes. Only fully covered viral genomes are reported. (**E**) Reads are assigned to assembled viral genomes. Read may be shared across two or more viral genomes. VGA infers relative abundances of viral genomes using the expectation–maximization algorithm. (**F**) Long conserved regions are detected and phased based on expression profiles. In this example red and green viral genome share a long conserved region (colored in black). There is no direct evidence how the viral sub-genomes across the conserved region should be connected. In this example four possible phasing are valid. VGA use the expression information of every sub-genome to resolve ambiguous phasing

different viral genomes if they share different SNVs in the overlapping region. Viral variants are identified from the graph as independent sets of non-conflicting reads. Non-continuous coverage of rare viral variants may limit assembly capacities, indicating that increase in coverage is required to increase the assembly accuracy. Frequencies of identified variants are then estimated using an expectation–maximization algorithm. Compared with existing approaches, we are able to detect rare population variants while achieving high assembly accuracy.

2.2 Error correction

The proposed sequencing is able to eliminate errors from sequencing data and produce highly accurate read sequences. It uses a high-fidelity sequencing protocol that attaches individual barcodes during the library preparation step. The barcodes are then used to identify reads originated from the same fragment, allowing to access multiple sequencing data of the same fragment. It follows that every sequenced position of the fragment would have multiple independent evidence, suitably promoting highly accurate consensus reads. By applying an error-correction procedure of the protocol, we are able to address both sequencing and PCR errors, which leads to high assembly accuracy.

2.3 Consensus construction

We build a consensus from paired-end reads using Vicuna (Yang *et al.*, 2012). Our sequencing method should not contain any particularly low coverage region allowing reconstruction of population consensus for viral sample. In the event that Vicuna produces multiple contigs rather than a complete consensus, we use BLAST to merge contigs. We require 50 nt overlap to merge any pair of contigs. In the next step, the population consensus is used as a reference genome to map reads. Building the reference genome from actual sequencing data rather than using an annotated genome provides us with an accurate and unique mapping.

2.4 Read mapping

As with many viral population analyses, the first step of VGA is to map the reads. We map reads onto the *de novo* consensus using InDelFixer (Armin and Beerenwinkel, 2013) with default parameters. False read alignments are filtered out using fragment length distribution inferred from the mapping data. Assuming that the fragment length follows a normal distribution (Hormozdiari *et al.*, 2009), we only keep reads with fragment length within three standard deviations from the mean. In total 1.2% of reads have been filtered out versus expected .3% according to the three-sigma rule.

2.5 Viral population assembly

The combination of deep coverage with high accuracy provides an unprecedented opportunity for estimating genomic diversity in a viral population. The viral population assembly starts with determining pairs of mapped reads conflicting with each other in the overlapping region. Following Huang *et al.* (2011), we construct the conflict graph $G=(V,E)$ with vertices corresponding paired-end reads, i.e. $V = R$, and edges connecting conflicting pairs of paired-end reads.

Obviously, any true viral genome corresponds to a maximal independent set in the conflict graph (i.e. a maximal set of pairwise nonadjacent vertices), although not every maximal independent set necessarily corresponds to a true viral genome. We adopt a parsimonious approach requiring to cover the conflict graph with the minimum number of maximal independent sets. This problem is equivalent to MIN-GRAPH-COLORING which is NP-hard. There exists many heuristics for solving this problem [see, e.g. Johnson and Trick, 1996; Kubale, 2004] based on greedy selection of a maximal independent set. Unfortunately, our attempts to build even a single viral genome failed, as it is difficult to arrange paired-end reads into a connected single path. Indeed, a greedy algorithm runs out of any possible extension after just a few steps while concatenating paired-end reads from left-to-right.

Instead, we apply an alternative 'top–down' approach of recursive graph partitioning along the maximum cut (Max Cut) which has been previously successfully applied for human haplotyping (Duitama *et al.*, 2012). Given a graph $G=(V,E)$, the Max Cut problem asks for partitioning of the vertices into two components $V=V_1 \cup V_2$ maximizing the total number of edges which have one endpoint in V_1 and the other in V_2. The Max Cut problem though NP-hard is well approximated by a simple 0.5-approximation algorithm that randomly assigns vertex to one of the two components (Mitzenmacher and Upfal, 2005). Our Max Cut heuristic starts with alternatively assigning left-to-right sorted mapped reads to two components and then repeatedly moves one vertex at a time from one component to another, improving the solution at each step, until no more improvements of this type can be made.

Our coloring heuristic recursively partitions the conflict graph until each component becomes independent. If reads of a given color completely cover the consensus genome, then the resulted sequence is accepted as the next viral genome. Otherwise, if assembled genome contains gaps, we add non-conflicting reads from other color classes in left-to-right order in attempt to fill the gaps. If all SNV positions are covered, then a newly reconstructed viral genome is added to the set $\mathcal{VG}$. Finally, the genomes whose gaps cannot be filled with the above procedure are dropped.

Algorithm 1: VGA Assembly Algorithm.

Input: Set of reads R aligned to the consensus genome
Build conflict graph G = (V,E) from set R
Recursively color G into color classes $\mathcal{C}$ using Max Cut
Initialize the set of complete viral genomes $\mathcal{VG} \leftarrow \varnothing$
for each color class $c_i \in \mathcal{C}$ **do**
  Compute maximal independent set in $G=(V,E)$ containing c_i
  Assemble reads in c_i into viral genome g_i
  if g_i covers all positions in the consensus genome **then**
    $\mathcal{VG} \leftarrow \mathcal{VG} \cup \{g_i\}$
  end if
end for
Output: Set of complete viral genomes $\mathcal{VG}$

2.6 Viral population quantification

In the final step of the workflow, an expectation–maximization algorithm is used to infer the relative abundances of assembled viral quasi-species similar to what is described in Eriksson *et al.* (2008). We extend the previous EM and likelihood formulation to incorporate a prior probability for the viral population and compute the *maximum a-posteriori* estimate, rather than the MLE.

Let H be a random variable over the set of viral variant genomes $\mathcal{H} = \mathcal{VG}$ and let R be a random variable over the set of reads $\mathcal{R}$. Let $p[\mathcal{H}] \sim \mathrm{Dir}(\alpha,\ldots,\alpha)$ be the prior probability of observing a given set of variants and denote $p_h = \Pr[H=h]$ to be the probability of observing a particular variant h. The probability of observing read $r \in \mathcal{R}$ is given by marginalizing over all variants

$$\Pr[R=r] = \sum_{h \in \mathcal{H}} \Pr[R=r|H=h] \cdot p_h$$

where

$$\Pr[R=r|H=h] = \begin{cases} 1/K_h & \text{if } r \text{ is consistent with } h \\ 0 & \text{otherwise} \end{cases}$$

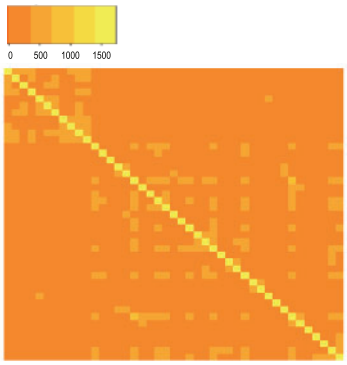

Fig. 3. Genomic architecture of 44 real HCV viral genomes from 1739-bp-long fragment of E1E2 region. Length of longest common region shared between any two viral genomes is represented by color

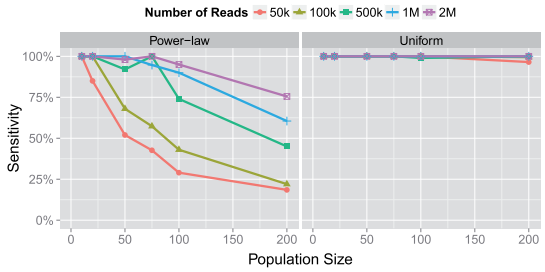

Fig. 4. Accuracy of population size prediction. Up to 200 viral genomes were generated from the Gag/Pol 3.4 kb HIV region. The population diversity is 5–10%. Viral genome abundances follow power-law and uniform distributions. Consensus error-corrected 1002 bp paired-end reads were simulated from HIV population

and K_h is the number of reads consistent with h. We can now define the log-posterior as

$$\log \Pr[\mathcal{H}|\mathcal{R}] = \sum_{r\in\mathcal{R}} n_r \cdot \log \Pr[R=r] + \alpha \cdot \sum_{h\in\mathcal{H}} \log p_h - C_{\mathcal{R}}$$

where $C_{\mathcal{R}}$ is a constant and n_r is the number of reads r. As this function is non-convex and difficult to optimize, we solve the easier problem of maximizing its lower-bound,

$$\sum_{r\in\mathcal{R}} \sum_{h\in\mathcal{H}} n_{rh} \cdot \log\left(\Pr[R=r|H=h] \cdot p_h\right) + \alpha \cdot \sum_{h\in\mathcal{H}} \log p_h$$

where n_{rh} is the expected number of reads r generated by variant h. The EM algorithm computes this by

$$n_{rh} = n_r \cdot \frac{p_h \cdot \Pr[R=r|H=h]}{\Pr[R=r]},$$

and subsequently maximizes the log-posterior with the MAP estimate given by

$$\hat{p_h} = \left(\alpha + \sum_{r\in\mathcal{R}} n_{rh}\right) / \left(\alpha + \sum_{r\in\mathcal{R}} n_r\right)$$

3 RESULTS

3.1 Performance of VGA on simulated data

Because the ground truth is unknown for sequenced viral populations, simulations present a standardized way to assess the performance of viral assembly tools. The proposed high-fidelity protocol allows to correct sequencing errors, thus giving access to highly accurate sequencing data. Post-sequencing error-correction techniques are available for reads obtained by regular protocol offering the possibility to partially correct sequencing errors trading off for real biological mutations. Grinder (Angly *et al.*, 2012) is used to generate reads from both the high-fidelity and regular sequencing protocol. Reads are generated from both real and synthetic viral variants with different sequencing parameters and viral expression profiles. Grinder is a state-of-the-art sequencing read simulator able to produce shotgun sequencing data from a viral population with different expression profiles. We mapped the simulated paired-end reads onto the consensus using Mosaik. The consensus was constructed using Vicuna (Yang *et al.*, 2012), a *de novo* assembly tool able to produce a linear consensus from deep paired-end sequencing data (see Section 2.3 for details).

We use sensitivity and positive predictive value (PPV) to evaluate the quality of viral genomes assembled by VGA. We consider fully assembled viral genome without errors. Sensitivity is defined as the portion of assembled quasi-species that match true quasi-species, i.e. $Sensitivity = TP/(TP+FN)$. Positive predictive value is defined as the portion of true sequences among assembled sequences, i.e., $PPV = TP/(TP+FP)$. Additionally, we evaluate ability of our method to estimate population size (i.e. number of viral genomes in the population). Accuracy of population size prediction is defined as a ratio between estimated and true population sizes. Finally, we use Jensen–Shannon divergence (JSD) to measure the accuracy of frequency estimation. Given two probability distributions, JSD measures the 'distance' between them, or in other words, the quality of approximation of one probability distribution by the other distribution. It is defined as the Kullback–Leibler divergence from distributions P and Q to their mixture. Formally, the JSD between true distribution P and approximation distribution Q is given by the formula

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{D}_{KL}(P||M) + \frac{1}{2}\mathrm{D}_{KL}(Q||M)$$

where Kullback–Leibler divergence D_{KL} is

$$\mathrm{D}_{KL}(P||Q) = \sum_{i=1}^{n} P(i)\log\frac{P(i)}{Q(i)}$$

and $M = \frac{1}{2}(P+Q)$. The motivation for using JSD is a consequence of KL divergence being undefined when assembly methods fail to reconstruct some variant i, hence forcing $Q(i)$ to be 0. JSD averts this by measuring the distance to the mixture, which contains all true and called variants (TP and FP).

Our first simulated study compares the assembly accuracy across different virus species. We focus on effect of read-length and throughput on assembly quality for different types of viruses. Paired-end reads of various length corresponding to high-fidelity and regular sequencing protocols are simulated from HIV and HCV populations assuming uniform and power-law distributions. A power-law distribution (i.e. frequency

Fig. 5. Assembly accuracy estimation. Up to 200 viral genomes were generated from the Gag/Pol 3.4 kb HIV region. The population diversity is 3–20%. Viral genome abundances follow power-law and uniform distributions. Consensus error-corrected 2100 bp paired-end reads were simulated from HIV population

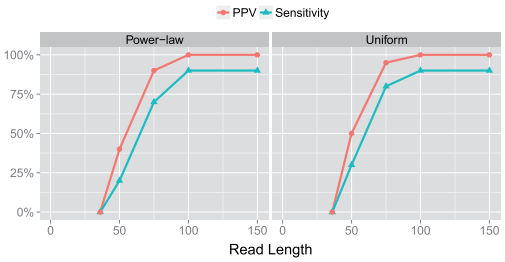

Fig. 6. Assembly accuracy estimation. Up to 200 viral genomes were generated from the Gag/Pol 3.4 kb HIV region. The population diversity is 3–20%. Viral genome abundances follow power-law and uniform distributions. Consensus error-corrected 2×100 bp paired-end reads were simulated from HIV population

of an individual viral genome is a power of the previous one) corresponds to a population with several dominant variants and many rare variants. The uniform distribution has equal frequencies for all viral genomes. HCV population is presented by 1739-bp-long fragment from the E1E2 region of 44 real HCV sequences. HIV population consist of 10 real intra-host viral variants mixture from 1.3-kb-long HIV-1 region, which included pol protease and part of the pol reverse transcriptase (Zagordi *et al*., 2010).

The genomic architecture across virus species was investigated and its influence on assembly accuracy was studied. HCV virus exhibits more complex genomic architecture with lower population diversity and longer conserved regions (Fig. 3) than HIV. Conserved regions were present in both viruses, although only HCV contains conserved regions longer then 450 bp. Conserved regions longer then the average fragment length (450 bp) may introduce ambiguity in the assembly process due to a lack of direct evidence of sub-genomes phasing across the conserved region. We performed simulated sequencing experiment where the average fragment amplification rate is 5, resulting in a five time decrease in throughput due to the consensus error correction performed by the high-fidelity sequencing protocol. Also the simulation experiments were adjusted to simulate a non-uniform amplification rate. Non-uniform amplification rate results in discarding fragments with insufficient amplification rate (<3). From

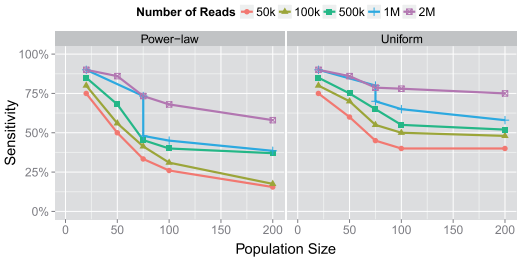

Fig. 7. Assembly accuracy estimation. Consensus error-corrected paired-end reads of various lengths were simulated from a mixture of 10 real viral clones from 1.3-kb-long HIV-1 region. Assembly accuracy as measured by PPV and sensitivity. Results are for 50 000 reads, no improvement was observed when increasing the number of reads

real studies it is known that around 10% of fragments are amplified less than three times. Sequencing errors produced by the regular protocol limited the ability of VGA to accurately assemble a viral population. All assembled variants contained large number of mismatches, additionally VGA significantly overestimated population size.

As expected, short read lengths dramatically inhibit reconstruction, which is evidenced by VGA failing to produce any full-length genomes when given 2×36 bp reads (Fig. 4). Because common regions for distinct HCV viral genomes are significantly longer than for HIV, it is not surprising that performance of VGA is worse on HCV data—for 3 M 2×150 bp reads simulated from 44 1739-bp-long viral genomes, sensitivity is 50%, PPV is 80%. Results on HCV data confirm that the lower mutation rate and presence of conserved regions have a negative impact on the ability to accurately reconstruct individual viral genomes. Surprisingly, increasing the read length for HIV from 100 to 150 bp yields no benefits for reconstruction accuracy suggesting that 100 bp read length is enough to distinguish between HIV viral variants with high mutation rate. Although further experiments are needed to determine optimum read length, our simulations suggest that 2×100 bp is recommended for small HIV viral populations and 2×150 bp is recommended for medium HCV population with complex genomic architecture.

We separately analyzed the ability of our method to estimate the viral population size (i.e. number of genomic variants present in the population). Non-continuous coverage limits the ability of the method to assemble full-length viral variants. To evaluate the accuracy of population size estimation, we compared the true population size known from simulated data with estimated results. Continuous coverage of each individual viral genome present in the sample has a strong impact on quality of population assembly. The probability of non-continuous coverage increases dramatically for viral genomes with low abundance. Thus, the presence of coverage gaps for rare variants introduces additional challenges in the assembly process, making rare genomes unreachable by assembly tools. The number of problematic genomes can be reduced by increasing sequencing depth; however, it does not guarantee complete elimination. While complete assembly of all such genomes is unrealistic, it is still possible to estimate the number of viral genomes present in the sample (population size). The number of independent sets reported by VGA provides us with an accurate population size estimation. Intuitively,

predicting the population size of a large viral population with many rare variants is more difficult than predicting for uniformly distributed or small populations (Fig. 5). The predicted population size may serve as an indication of insufficient coverage to detect the full viral diversity present in the sample.

Deep coverage is a key for accurate estimation of underlying viral diversity. One such platform capable of offering millions of sequencing reads is Illumina HiSeq. The relatively short length of the produced reads is compensated for by sequencing the same fragment from both ends; therefore, producing coupled reads separated by a 'gap', known as paired-end read. To our knowledge, VGA is the first method scalable to millions of short paired-end sequencing reads able to produce full-length viral variants spanning the entire viral genome. We explore the influence of sequencing depth on the reconstruction accuracy for varying population structures (uniform and power-law distributions of viral genomes within the population). HIV-1 is known to have greater genetic variability than any other known virus (Ndungu and Weiss, 2012). The diversity among viral genomes in an HIV population can vary from 3 to 20% depending on regions (Martins *et al.*, 1992; Yoshimura *et al.*, 1996). Heterogeneous viral samples were prepared by generating viral populations from the Gag/Pol 3.4 kb HIV region. We simulated variant abundances adhering to either a uniform or power-law distribution. Not surprisingly, our simulations suggest that increased sequencing depth has a direct positive effect on the discovery of rare variants and improves the overall assembly accuracy. Figure 6 shows the effect of coverage and population size on assembly for reads of length 100 bp. Throughout all experiments, VGA maintained a PPV value of 100%.

In addition to point mutations, genetic recombination facilitates rapid evolution and production of diverse HIV genomes. Indeed, co-infected cells may produce recombinant viral progeny at levels lower than mutation rates in an intra-patient environment (Neher and Leitner, 2010). Hence, simulated datasets must account for both possible phenomena when determining the quality of assembly. We utilize a simulation model able to integrate both point mutations and recombination in the generated viral population depending on the amount of diversity required. A mixture of 10 real intra-host viral variants from 1.3-kb-long HIV-1 form the basis population. In addition to point mutations, our simulation model implicitly produces recombinant genomes by first constructing the genotype (i.e. sequence of SNVs) for the population. A random walk is performed over this genotype as specified number of times. Any cross-over that occurs represents a new recombination between the 'left' and 'right' original genomes. Recombinations are implicitly produced, and no control is imposed over number and length of the recombination. This model produces highly recombinant data on average, posing challenges for assembly and can be used to assess assembly quality. Simulation model incorporate mutation into the process by selecting a position and nucleotide-swap uniformly at random. Simulation results (Fig. 7) suggest that our method can accurately assemble viral population in presence of recombinations and point mutations, maintaining PPV of 100%.

Finally, we evaluate population quantification accuracy, i.e. the accuracy of our method in predicting abundances of the assembled variants. Taking the results from VGA on 10 real HIV clones with 50 000 reads and 2×100 bp, the JSD was 2.93e-05 for the Power-law and 0.001 for the Uniform-based populations. This already small measure only decreases as the size of the input grows.

3.2 Performance of existing viral assemblers on simulated consensus error-corrected reads

We have evaluated the performance of ShoRAH (Zagordi *et al.*, 2011) and QuasiRecomb (Zagordi *et al.*, 2012) for simulated consensus error-corrected read data.

ShoRAH disregards pairing information of reads, but it is scalable enough to handle up to 1 M reads. ShoRAH fails to produce full-length viral genome but reliably spans 98% of the consensus genome. It reasonably estimates the number of different viral genome, but even the most accurate ShoRAH-assembled viral genome differs from the closest true 1.3-kb-long viral genome in five nucleotides.

QuasiRecomb is designed to handle paired-end read data and manages to produce full-length viral genomes. Unfortunately it can reliably process no more than 100 K reads. Also the number of assembled distinct viral genomes is 10–200 times more than the number of true distinct viral genomes. The most accurate QuasiRecomb-assembled viral genome still differs from the closest true 1.3-kb-long viral genome in four nucleotides.

Unfortunately we could not compare our method with QColors (Huang *et al.*, 2011) assembly algorithm, which uses a similar conflict graph to represent viral population. A CSP solver is used by QColors for coloring the graph which may limit its scalability to high-throughput datasets consisting of millions of sequencing reads. Currently, QColors is not publicly available (Upon querying for information on obtaining QColors, the authors were informed that the original software was tightly coupled for the analyses done in its original manuscript, and is not currently available for general use.).

3.3 Performance of VGA on real HIV data

To further test the ability of VGA to accurately assemble a diverse natural occurring population and predict variant abundance levels, we used an Illumina HiSeq HIV dataset, which consisted of 15 M 2×100 bp paired-end reads with attached barcodes. Next, the high-fidelity sequencing protocol able to eliminate sequencing errors was applied resulting in 3.2 M consensus error-corrected reads (further referred to as reads). The reads were then used to build *de novo* population consensus using Vicuna 1.3 (Yang *et al.*, 2012). When run on our real data, Vicuna produced four contigs of average length 1195 bp. Each contig was then run through BLAST to check for overlaps. Once overlaps were found, the contigs were assembled into a final consensus of length 4337 bp.

Validation of *de novo* consensus. A *de novo* assembled consensus was compared against reference-based consensus. To produce reference-based consensus, we iteratively map reads onto the HIV reference (Gag/Pol 3.4 kb HIV region) using InDelFixer. InDelFixer iteratively changes the reference genome based on the mapping of the current iteration. Also, we used InDelFixer in single iteration mode to map reads onto the constructed *de novo* consensus. *De novo* consensus is longer (4337 bp) than the reference-based consensus (3440 bp) and contained two regions with extremely peaked coverage compared with the surrounding

regions. Both regions were considered to be the result of technical artifacts and removed from further consideration. After removing both regions, the length of new *de novo* consensus becomes 3452 bp. We also filtered reads that belonged to regions with extreme coverage. Finally we compared the number of reads mapped to the reference-based consensus versus the *de novo* consensus. A larger amount of reads mapped to the assembled consensus, thereby highlighting the advantage of *de novo* procedure for consensus construction over a reference-based.

From the *de novo* consensus, VGA assembled 32 full-length viral genomes that differ from each other in 2145 SNVs. Among known HIV sequences, Gag/Pol is the closest to the *de novo* consensus. Each of the 32 full-length viral genomes do not contain stop codons inside two known coding regions of Gag/Pol of length 1520 and 1820 bp, respectively. Alternatively, when VGA is applied to all 15 M original uncorrected reads, 57 distinct viral genomes are assembled among which 36 contain stop codons in the two coding regions. This shows that a regular sequencing protocol is unsuitable for viral genome reconstruction.

4 DISCUSSION

We have presented VGA, an accurate method for viral population assembly from ultra-deep sequencing data. The proposed algorithm is coupled with a high-fidelity sequencing protocol able to eliminate errors from sequencing data. Deep coverage in combination with highly accurate data allows our method to accurately estimate the underlying diversity of a viral population. In particular, it makes possible to distinguish true biological mutations from sequencing errors, facilitating assembly of rare individual genomes. Our method condenses the viral population into a conflict graph built from aligned reads. To distinguish between viral variants, the conflict graph is colored into a minimal set of colors. Each color represents individual viral genomes composed from the set of non-conflicting reads. An expectation-maximization algorithm was used to estimate relative abundance frequencies of assembled viral genomes.

To our knowledge, our method is the first viral assembly method that scales to millions of paired-end sequencing reads. Experiments on both real and synthetic HIV datasets generated with various sequencing parameters and distribution assumptions suggest that VGA is able to assemble diverse viral population from millions of paired-end reads. The ability of our method to maintain 100% assembly accuracy makes it suitable for clinical applications. In addition, the constant increase of sequencing depth offered by high-throughput technologies provide us with unprecedented resolution promising to increase number of discovered ultra-rare viral variants in the population.

ACKNOWLEDGEMENTS

The authors thank three anonymous reviewers for helpful comments, UCLA Clinical Microarray Core for performing the high-throughput sequencing experiment.

Funding: S.M. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448 and 1320589, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01- HL30568, P01-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. S.M. was supported in part by Institute for Quantitative & Computational Biosciences Fellowship, UCLA. N.C.W. was supported by Molecular Biology Whitcome Pre-Doctoral Fellowship, UCLA. A.Z. was partially supported by Agriculture and Food Research Initiative Competitive Grant no. 201167016-30331 from the USDA National Institute of Food and Agriculture and NSF award IIS-0916401. N.M. was partially supported by Second Century Initiative, Georgia State University.

Conflict of Interest: none declared.

REFERENCES

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94–e94.

Armin, T. and Beerenwinkel,N. (2013) http://www.bsse.ethz.ch/cbg/software/InDelFixer.

Astrovskaya,I. (2011) Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, **12** (**Suppl. 6**), S1.

Bansal,V. and Bafna,V. (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, i153–i159.

Duitama,J. *et al.* (2012) Fosmid-based whole genome haplotyping of a hapmap trio child: evaluation of single individual haplotyping techniques. *Nucleic Acids Res.*, **40**, 2041–2053.

Eriksson,N. *et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, **4**, e1000074.

Gnerre,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Henn,M.R. *et al.* (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.*, **8**, e1002529.

Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.

Huang,A. *et al.* (2011) QColors: an algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. *In Silico Biol.*, **11**, 193–201.

Johnson,D.S. and Trick,M.A. (1996) *Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge, October 11-13, 1993*. Vol. 26, American Mathematical Society, USA.

Kinde,I. *et al.* (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA*, **108**, 9530–9535.

Kubale,M. (2004) *Graph Colorings*. Vol. 352, American Mathematical Society, USA.

Lauring,A.S. and Andino,R. (2010) Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.*, **6**, e1001005.

Liu,J. *et al.* (2011) Analysis of low-frequency mutations associated with drug resistance to raltegravir before antiretroviral treatment. *Antimicrob. Agents Chemother.*, **55**, 1114–1119.

Luo,R. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, **1**, 18.

Mancuso,N. *et al.* (2011) Reconstructing viral quasispecies from NGS amplicon reads. *In Silico Biol.*, **11**, 237–249.

Martins,L.P. *et al.* (1992) Complex intrapatient sequence variation in the V1 and V2 hypervariable regions of the HIV-1 gp120 envelope sequence. *Virology*, **191**, 837–845.

Metzker,M.L. (2009) Sequencing technologiesthe next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Mitzenmacher,M. and Eli,U. (2005) *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge.

Ndungu,T. and Weiss,R.A. (2012) On HIV diversity. *AIDS*, **26**, 1255–1260.

Neher,R.A. and Leitner,T. (2010) Recombination rate and selection strength in hiv intra-patient evolution. *PLoS Comput. Biol.*, **6**, e1000660.

Palmer,S. *et al.* (2006) Selection and persistence of non-nucleoside reverse transcriptase inhibitor-resistant HIV-1 in patients starting and stopping non-nucleoside therapy. *AIDS*, **20**, 701–710.

Prosperi,M.C. and Salemi,M. (2012) QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, **28**, 132–133.

Tsibris,A.M. *et al*. (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy *in vivo*. *PLoS One*, **4**, e5683.

Wang,C. *et al*. (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*., **17**, 1195–1201.

Yang,W.-Y. *et al*. (2013) Leveraging multi-SNP reads from sequencing data for haplotype inference. *Bioinformatics*, **29**, 2245–2252.

Yang,X. *et al*. (2012) *De novo* assembly of highly diverse viral populations. *BMC Genomics*, **13**, 475.

Yoshimura,F.K. *et al*. (1996) Intrapatient sequence variation of the gag gene of human immunodeficiency virus type 1 plasma virions. *J. Virol*., **70**, 8879–8887.

Zagordi,O. *et al*. (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, **12**, 119.

Zagordi,O. *et al*. (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol*., **17**, 417–428.

Zagordi,O. *et al*. (2012) Probabilistic inference of viral quasispecies subject to recombination. In: *Research in Computational Molecular Biology*. Springer Berlin Heidelberg, pp. 342–354.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*., **18**, 821–829.