VirGenA: a reference-based assembler for variable viral genomes

Gennady G. Fedonin, Yury S. Fantin, Alexnader V. Favorov,
German A. Shipulin and Alexey D. Neverov

Corresponding author: Gennady G. Fedonin, 111123, Novogireevskaya st., 3a, Moscow, Russia. Tel.: +7 (903) 232 14 41; E-mail: gennady.fedonin@gmail.com

Abstract

Characterization of the within-host genetic diversity of viral pathogens is required for selection of effective treatment of some important viral infections, e.g. HIV, HBV and HCV. Despite the technical ability of detection, there are conflicting data regarding the clinical significance of low-frequency variants, partially because of the difficulty of their distinguishing from experimental artifacts. The issue of cross-contamination is relevant for all highly sensitive techniques, including deep sequencing: even trace contamination leads to a significant increase of false positives in identified SNVs. Determination of infections by multiple genotypes of some viruses, the incidence of which can be considerable, especially in risk groups, is also clinically significant in some cases. We developed a new viral reference-guided assembler, VirGenA, that can separate mixtures of strains of different intraspecies genetic groups (genotypes, subtypes, clades, etc.) and assemble a separate consensus sequence for each group in a mixture. It produced long assemblies for mixture components of extremely low frequencies (<1%) allowing detection of cross-contamination of samples by divergent genotypes. We tested VirGenA on both clinical and simulated data. On both types of data, VirGenA shows better or similar results than the existing *de novo* assemblers. Cross-platform implementation (including source code) is freely available at <https://github.com/gFedonin/VirGenA/releases>.

Key words: viral genome assembly; NGS; genotyping; mixture separation

Introduction

Characterization of the within-host genetic diversity of viral pathogens is required for selection of effective treatment of some important viral infections, e.g. HIV [1], HBV [2] and HCV [3]. Genome-wide deep sequencing can be used to detect low-frequency variants (having frequencies <1%) resistant to treatments in such heterogeneous viral populations. However, there are conflicting data regarding the clinical significance of detection of low-frequency variants of HIV-1 [4–6]. Association of mutations with the outcome of HIV treatment for non-nucleoside

reverse-transcriptase inhibitors is not confirmed in many studies [7, 8]. False-positive variations can be one possible reason for contradictions between studies of associations of low-frequency mutations with success of therapy [8]. The problem of cross-contamination between samples is relevant to all high-sensitivity techniques, including deep sequencing [9, 10]. Contamination of the sample by another sample, even in a small proportion, leads to a significant increase of false positives in identified SNVs [9]. We do not know an automatic method to detect the presence of cross-contamination between samples of highly variable virus isolates. To determine the

Gennady Gennadievich Fedonin is a bioinformatics programmer in the bioinformatics group in the Department of Molecular Diagnostics at Central Research Institute for Epidemiology and a researcher in the Institute for Information Transmission Problems of the Russian Academy of Sciences.

Yury Sergeyevich Fantin is a software developer in the Department of Molecular Diagnostics at Central Research Institute for Epidemiology.

Alexnader Vladimirovich Favorov is a research associate at the Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, the Johns Hopkins University and a senior research fellow at the laboratory of systems biology and computational genetics at the Vavilov Institute of General Genetics of the Russian Academy of Sciences and at the laboratory of bioinformatics of the Research Institute of Genetics and Selection of Industrial Microorganisms.

German Alexandrovich Shipulin is the head of the Department of Molecular Diagnostics at Central Research Institute for Epidemiology.

Alexey Dmitrievich Neverov is the leader of the bioinformatics group in the Department of Molecular Diagnostics at Central Research Institute for Epidemiology.

contamination of a sample of HIV isolates by other samples, Zanini et al. [11] built the distribution of similarity of reads with the consensus sequence and discarded reads with extremely low similarity. The disadvantage of this method, besides that its reliance on subjective evaluation of diversity, is the inability to distinguish contamination from infections by multiple genotypes, the frequency of which can be considerable, especially in risk groups: for HIV ~10% of the population of HIV-1 positive [12–14], for hepatitis C 5–25% [15, 16] and for hepatitis B 5–37% [17, 18]. Determination of mixed genotypes is clinically significant in some cases, such as hepatitis C virus [19].

Mapping reads on the reference sequence historically has been the first method used to assemble the viral genomes [20–22]. At the same time, for highly variable viruses like HIV using a distant reference may cause variant calling errors because of misalignment of reads or incomplete assembly of genome fragments of high divergence [20, 23]. Using any fixed reference may be problematic, as viral consensus may differ significantly between patients, and it may be difficult to choose a suitable reference from any sequence database, which allows obtaining high-quality alignments of reads [24, 25]. Another solution is first to assemble consensus sequence *de novo*, and then use it as a reference for the reads mapping and detection of variants [24]. For this, specially designed viral *de novo* assemblers, such as Vicuna [26] and IVA [27], were proposed.

We have developed a new reference-guided assembler of highly variable viral genomes, VirGenA, based on iterative mapping and *de novo* reassembling of highly variable regions, where similarities of mapping reads to a specified reference are too low. The method can handle distant reference sequence because of a specially designed read mapper. In cases when a good reference is unknown, the best performance is achieved by using a user-provided, multiple aligned reference sequences as vocabulary.

We compared assemblies produced by VirGenA and two programs for *de novo* assembly: Vicuna [26] and IVA [27] on simulated and clinical HIV-1 data sets and on the clinical HCV data set. For this purpose, we developed a method for comparing different algorithms of assembly of highly variable viral genomes, which is based on the assumption that a virus establishes a heterogeneous population of haplotypes—quasispecies within the host.

One of the advantages of VirGenA over other programs is the possibility of genotyping, e.g. of determining the isolate's membership in genetic groups (genotypes; subtypes; clades), on which genetic classification within viral species is based. Our method allows us to genotype samples comprising more than one subpopulation of the virus, each of which pertains to a different genetic group. Prior genotyping is required for correct interpretation of some SNVs that are clinically significant on a particular genetic background and benign on others [28, 29]. SNV analysis greatly benefits if reads are sorted and assembled according to corresponding genetics groups that are presented in a particular sample in comparison with the traditional approach where contigs correspond only to the dominating genetic group used for mapping of a pool of reads. In the latter case, some of the detected SNVs correspond to differences between subtypes, whereas some SNVs are artifacts of approaching of multiple alignment by mapping procedure (Figure 1).

Some mixtures may be artifacts of cross-sample contamination. With the help of VirGenA, we find the evidence of cross-contamination for some sequenced HIV-1 clinical samples from the data, which is publicly available from the Sanger Institute

(<http://www.ebi.ac.uk/ena/data/view/PRJEB2262>), and for some samples from the HCV data set [30].

Material and methods

Data sets

We evaluated performance of our algorithm in comparison with Vicuna [26] and IVA [27] on two data sets of HIV and HCV clinical samples. HIV data set consists of 164 HIV samples sequenced on Illumina platform (250bp paired reads) by the Sanger institute (<http://www.sanger.ac.uk/resources/downloads/viruses/retroviridae-lentivirus.html>) available at <http://www.ebi.ac.uk/ena/data/view/PRJEB2262>. Some of them were previously used for assessment of IVA accuracy [27].

The HCV data set consists of 22 clinical samples that were single component and of 16 samples that were mixtures of two subtypes according to genotyping in reference laboratories or were artificial mixtures [30]. All samples are sequenced on Illumina platform (100bp paired reads). For each sample, there were two sets of reads corresponding to two different protocols of sequencing libraries preparation: metagenomic without enrichment (O-Meta) and enrichment by a capture on hybridization probes (O-Capt). As publicly available FASTQ files contain no information about the protocol used for preparation of a particular sample, we classified them ourselves based on the expectation that for the O-Meta sequencing libraries, fractions of nonhuman reads mapped on HCV multiple sequence alignment (MSA) would be sufficiently lower than for O-Capt (see Supplementary Materials, 1.1 for details). Results for O-Capt HCV data set are presented in the main text of the manuscript; results for O-Meta data set are in the Supplementary Materials Sections 2.2 and 2.3.

Algorithm

Our approach (Supplementary Figure S1) is based on a simple idea: if a good reference sequence is available, we use iterative mapping of reads to obtain a corresponding consensus sequence; otherwise, if an MSA of references, which represent viral diversity, is available, we group reads into subpopulations seeded by reference sequences from the alignment, and then assemble consensus for each population separately in the same way as in the single-reference case. Thus, we evade dealing with calling and interpretation of lineage-specific mutations and artifact SNVs (Figure 1). The procedure for the search of the most suitable references produces a minimal subset of references most similar to a pool of reads than each pair of reads ascribed to one or more references depending on the joint score of alignments. Finally, we postprocess results in the case of using of reference MSA as a vocabulary to guide assembly: we conservatively assume that long fragments of high similarity (>99%) in contigs assembled following different references are made from reads of dominating subpopulation and cut such fragments out from contigs with lower coverage.

Data preprocessing

Adapter and polymerase chain reaction primer sequences (if available) should be deleted before start of assembly. We used Trimmomatic [31] for trimming both adapter and primer sequences (see Supplementary Materials, 1.2, for details).

We use the hashing algorithm for reads' mapping: first, we quickly localize each read on a reference genome using hash table storing all short k -mer positions, and then, we use the Smith–Waterman algorithm to precisely align a read to the genome at a location with the highest score. We generalize the read localization approach for MSA, using the hash table of alignment positions of all possible k -mers in reference genomes. To distinguish a correct localization of a read from a spurious localization, we use a threshold on read localization score. The threshold depends on read length and on the specified level of significance (0.01) that has been estimated on simulated reference genomes. In general, our read mapping approach is similar to that of MOSAIK [32], although they are not completely similar (see Supplementary Materials. 1.3, for details).

Our approach can assemble consensus sequence by mapping reads either on a user-specified reference sequence or, if an MSA of reference sequences is provided, on a minimal subset of references most similar to the sequencing data. Here, we briefly describe the procedure of selection of the optimal subset of references. First, we localize all reads on reference MSA and cluster adjacently localized reads with each other using Usearch [33] with given similarity threshold (95%); then for each cluster, we build a consensus sequence (short contig). For each short contig, we select a subset of references by picking the reference having the highest local similarity (Sim_{max}) and all references having

similarity $\text{Sim} > \text{Sim}_{\max} - \delta$, where δ is a specified threshold (5%). Finally, we build long contigs by traversing the overlap graph built on short contigs, where selected references are used as constraints on possible paths, providing information about long-range linkage of short contigs. A similar idea was previously used for analysis of mixed chromatograms for the direct Sanger sequencing method [34]. The graph traversing algorithm links the unvisited short contigs, which have common references in their reference sets into haplotypes, and then penalizes already visited contigs at the subsequent iterations. To do that, an initial positive weight is assigned to each short contig. The path traversing algorithm may use one short contig in multiple paths, but if a short contig has already been used, its weight is assigned to zero; thus, the path traversing algorithm at each step prefers to track through unvisited contigs. Each path through overlap graph corresponds to a long contig. We use the following greedy strategy to obtain a minimal number of references from the reference set, so that the references are most similar to our set of long contigs. For each reference from the minimal set, there is at least one long contig so that identity of the contig's pairwise alignment with the reference is less than the highest identity of alignment of that contig over all references from the reference set, but the difference is no more than a specified threshold. This threshold for our reference selection procedure is supposed to be set close or equal to the mean pairwise distance between viral genotypes that were used for genetic classification of the virus, e.g. for HIV-1 and HCV assemblies we use 5% threshold. All reads are then assigned to

one or more of the selected references (see Supplementary Materials, 1.4 for details); reads assigned to each reference are assembled by iterative mapping procedure described below; the overlap graph and long contigs are not used further in data processing.

If reference selection is disabled by the user, the VirGenA uses the default reference (e.g. HXB2 for HIV-1 and 4s.JF735136 for HCV). Also, the default reference is used for assembly in the rare cases, when no reference was selected by the procedure described above.

We select suitable references for assembling of HIV samples by our VirGenA program from the set of 332 aligned HIV-1 reference sequences, which represent most circulated viral subtypes and recombinant forms from NCBI Viral Genotyping Tool HIV-1 2009 RefSet [35]. For assembling of HCV samples, we used MSA of 174 reference sequences from NCBI Viral Genotyping Tool HCV RefSet [35].

Reference-guided assembly

After the references are selected, the *k*-mer-based similarity score is computed for each pair of a read and one of the selected references. Then, each paired read is assigned to one or more references that have the highest score with this read. Reads that assigned to a particular reference are then treated separately from other reads. All those reads are mapped on their corresponding references, and then a consensus sequence is built for an MSA of reads. We then repeat read mapping and consensus generation using the new consensus sequence obtained at the previous stage as a reference. Next, we cut out poorly aligned regions of MSA and close the arising gaps by *de novo* reassembly. To do this, we increment the edges of cuts step by step by one nucleotide at a time using reads that are partially mapped to the edges; the method is similar to that in [36] (see Supplementary Materials, 1.5, for details). After reassembling, all reads are mapped back to the assembly, and the final consensus sequence is generated (see Supplementary Materials, 1.5, for details).

Postprocessing

The assemblies may sometimes be chimeras: the contigs assembled on different references for one sample may contain identical or similar subsequences in the variable genome regions. Such subsequences are found by the nucleotide BLAST [37] of each contig against all the contigs assembled on other references and identifying BLAST high-scoring segment pairs (HSPs) with identity >99% threshold (for HIV1). Then, we identify all regions of references having local pairwise identities lower than the threshold (99%) and corresponding to nearly identical subsequences (identity >99%) in some contig pairs. Then, each of these subsequences is cut out from the contig that correspond to the reference, which has less assigned reads, and all the reads that have been aligned to the cut out subsequence are then realigned on the assembly that was built using the reference, which has more assigned reads. Finally, the new consensus sequences are computed. The reads aligned to each of the selected references are saved in separate BAM files (see Supplementary Materials, 1.6, for details). We also provide a converter (see Supplementary Materials, 1.7, for details) from VirGenA contigs output to SSPACE [38] input format for further scaffolding.

Comparing different algorithms for assembling highly variable viral genomes

As we did not have a gold standard for our data sets of clinical samples, we compared the results of different assemblers with the set of reference genomes obtained from the curated databases that was used in our approach for HIV (<https://www.hiv.lanl.gov/content/sequence/NEUALIGN/align.html>) and HCV (<https://hcv.lanl.gov/content/sequence/NEUALIGN/align.html>) assembling. For the maximal robustness, we excluded the sequences, which were used in corresponding vocabularies by VirGenA, from these reference sets. For each sample, we clustered all the contigs that were assembled by VirGenA, Vicuna and IVA, using CD-HIT with a threshold of 99% for local alignment identities and a threshold of 50% for portions of shorter sequences to be covered by alignments. Then, for each cluster, we selected the 'cluster reference set', i.e. a subset of references that have the BLASTN HSP normalized scores, NS ($NS = s/(m \cdot l)$), where s is raw BLASTN score, m is BLASTN match score and l is contig length) with a centroid sequence of the cluster (see CD-HIT manual for definition of centroid sequence) within the range of 5% below the highest HSP score. Finally, for each sample, we constructed a minimal reference set by the following greedy approach: at each iteration, the reference that is contained within the largest number of cluster reference sets was selected and included into the minimal set. If there were more than one such reference, we picked the one that had the highest average identity with the centroids of corresponding clusters. All clusters that included the candidate reference in their reference sets were not further considered, and all their contigs were assigned to the reference included in the minimal reference set. The process stopped when all clusters were assigned to references. References that were elements of a sample minimal reference set were considered as proxies of the unknown components of mixtures.

The same procedure was performed for the comparison of assemblers on simulated mixtures. For each assembly, we assigned each contig to a reference using almost the same procedure as described above. The difference was that for each artificial sample, we chose references from a set of two known virtual components of a mixture. In the case of MIX data set, we used the consensus sequences that were assembled by VirGenA for corresponding clinical samples, as references.

Further, we directly compared programs by pairwise identities between assemblies of sequencing reads obtained for each sample. Identity between two assemblies was calculated as follows: for each contig of the first assembly, we performed a BLASTN best HSP search in the second assembly, and then we calculated averaged identities of obtained best HSPs taking each identity with weight equal to the proportion of contig length to the total length of assembly. To make the value independent of query and subject selection, we used each assembly as a BLASTN query against the other assembly and calculated the minimum between two resulting identities (see Supplementary Materials, 1.9, for details).

Results

Comparison of performances of assemblers on simulated single-component samples

At first, we tested VirGenA and the other algorithms on a data set of 46 simulated single-component samples; for each sample, a FASTQ file containing 500 000 paired reads of length of 100 bp

with insertion length of 500bp was generated by pIRS [39] using HIV references that were absent in the HIV-1 dictionary used for assembling and had minimal identity to the closest reference in the dictionary. In total, 34 of 46 sequences were recombinants of two or more parental subtypes. Sequences for generating reads are the least similar to sequences in VirGenA vocabulary reference set: they were obtained by the same procedure as used for PIRS_min (see Supplementary Materials, 1.1, for details) for all subtypes present in NCBI Viral Genotyping Tool HIV-1 2009 RefSet (Supplementary Table S1).

The VirGenA algorithm selected single reference for each sample, proving that VirGenA in general did not predict false mixture components for single-component samples, and subtypes of selected references agreed with subtypes of references used for simulating the data in all but two cases. For the reference sequences (subtypes 29_BF and 16_A2D), used for simulation of reads of the discordant samples, we performed BLASTN search in VirGenA HIV-1 dictionary and found that both had HSPs with almost the same identities (91 and 90%, respectively) with two reference sequences. For both sequences, one of HSPs (selected by VirGenA) was of one of the parental subtypes (B and A2, respectively) and another was of the subtypes 29_BF and 16_A2D; however, query coverage was higher for the HSP with the sequences of subtypes B and A2.

VirGenA and IVA produced contigs of almost genome length (from 8000 to 10000) for 98 and 100% of the samples, respectively, significantly outperforming Vicuna (72%). Reference coverage by all algorithms was 100% for 32 samples; for 14 samples, VirGenA covered from 94 to 99% of the genome size, while the other assemblers covered the full genome. For all samples, identities were >99% for all methods (Supplementary Table S2).

Treating contamination by nontarget DNA

Some (or all) clinical samples used in this study might be contaminated by host DNA. Our algorithm can natively handle host genome contamination. Vicuna also has an option, which allows separating viral from the host reads based on multiple alignment of reference sequences of viral genome. The usage of IVA requires preliminary filtering of the host reads by mapping on the known host genome sequence. To filter out human reads from our data sets, we mapped reads to the hg19 by bowtie2 aligner [40] with default parameters. Assembled contigs from the HIV data set were classified as HIV if their best BLAST hits on HIV + hg19 nucleotide database were HIV sequences, while contigs from the HCV data set were searched against the VirGenA HCV reference sequence vocabulary keeping those that have best BLAST hit *E*-value lower than 10^{-10} . In further tests, we considered only target contigs (as defined above) produced by all algorithms on bowtie-filtered data; for Vicuna, we switched filtering option on using HIV alignment provided in Vicuna distribution and VirGenA HCV reference sequence vocabulary.

VirGenA performance on HIV clinical samples

We applied our VirGenA program to 164 HIV samples, with single reference selected for each of 142 samples (Supplementary Table S3), two references of different HIV-1 subtypes selected for 18 samples (Supplementary Table S4) and three for 1 sample (Supplementary Table S5); then for each sample assembly guided by selected references was performed. Three samples (ERR732125, ERR862194 and ERR862219) were assembled using default reference (HXB2) because not enough reads were mapped on our HIV reference alignment to perform reference

selection; *de novo* assemblies of these samples performed by Vicuna and IVA did not contain HIV contigs. In all of the 19 multicomponent mixtures, minor components comprised a low fraction (16 about <0.5%, one about 1% and one about 2.7%); however, even for such low admixtures, the VirGenA assemblies covered from 38 to 95% (median 73%) of reference length (Supplementary Tables S4 and S5). Major components of two-component mixtures covered from 58 to 100% (median 99%) of length of corresponding selected references. The assembled components in all but one of mixtures had higher similarity to corresponding reference sequences than cross-component similarities. Assembled component identities to their respective references were within estimated HIV intrasubtype similarity range (88–96% <http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/nomenclature/Nomen.html>) (Supplementary Tables S6 and S7). As all the multicomponent samples had low frequencies of minor components, we suspected cross-contamination for most of the mixed HIV samples.

To further investigate multicomponent samples, we clustered all relatively large contigs (length > 1000 bp), assembled by VirGenA for all samples, using CD-HIT [41, 42] with threshold of 99% for local alignment identities and a threshold of 50% for portions of shorter sequences to be covered by alignments (Supplementary Table S8). Totally, we obtained 169 clusters; 20 of them contained contigs from two or more samples, and 6 of these 20 clusters contained contigs from major components of some samples (or contigs from single-component samples) and contigs from minor components of two- and three-component mixtures. About 42% (8 of 19) of the multicomponent samples had minor components clustered with major components of the other samples. High similarity between the major components of some samples and the minor components of other samples also supported our cross-sample contamination hypothesis (see Supplementary Materials, 2.1, for details). Finally, we retrieved instrument IDs and multiplexing indices from the FASTQ files; the ID investigation showed that four of the six clusters, which were mentioned above, were produced on the same device. Furthermore, samples from these four clusters having the same device ID also had close (frequently successive) multiplexing indices. Therefore, these samples were produced on the same sequencing device and, probably, were obtained during the same experiment in one sequencing run, which supported our suggestion about cross-sample contamination.

Comparison of performances of assemblers on clinical data

HIV data set

For each sample, we constructed a minimal reference set and assigned contigs to corresponding references (see 'Methods' section). For each program and for each reference from a sample minimal reference set, we picked the longest assigned contig (Figure 2A). If for a particular reference no contigs were assembled by a particular program, we assigned a zero-length contig to that reference.

Most references in minimal reference sets of HIV samples (~74%) were picked exclusively by contigs assembled by Vicuna and having length of about 500 bp (two paired reads assembled together) (Figure 3A). The majority of assemblies made by Vicuna and VirGenA (63 and 77% correspondingly) contained nearly full genome-size ($8500 < \text{length} < 10000$) contigs, in contrast to IVA, with longest contigs for most samples (58%) of about half size ($4000 < \text{length} < 5000$) of the HIV-1 genome (Supplementary Table S9, see also Supplementary Figure S2).

We also compared the coverages of the reference sequences by aligned contigs assembled by assessed programs (Figure 2B and Supplementary Figure S3). The reference sequences used for test were not used for the assembling procedure (see ‘Methods’ section). First, we selected one reference sequence that have got maximal total coverage by contigs assembled by all assessed programs for each sample. For this sequence, all programs demonstrated similar performance, with VirGenA showing slightly superior results (Figure 2B). However, when the entire minimal reference sets were considered, we could see again that most of low-covered references (<14%) were covered exclusively by alignments with contigs assembled by Vicuna (Supplementary Figure S2). Vicuna assemblies had a lower (90.6%) [compared with IVA (92.0%) and VirGenA (92.1%)] median of distribution of average identities of local contig alignments with corresponding references across all samples and all references in corresponding minimal reference sets (Supplementary Figure S3). Results were similar if only one reference with maximal total coverage was considered for each sample (Supplementary Figure S4): Vicuna assemblies had higher identities on this data set (91.5%), but still slightly lower compared with IVA (92.0%) and VirGenA (92.2%). Thus, Vicuna assembled many short contigs having low identities with corresponding references.

Median pairwise identities across samples were high for IVA versus VirGenA (99.3%) (Supplementary Table S10; Figure 2C and D); in comparisons of Vicuna with VirGenA and IVA, median identities across samples were lower: 97 and 97% correspondingly. The analysis of cross-assembly similarities was done on the data set of all but three HIV-1 samples (ERR862219, ERR732125 and ERR862194): all programs except VirGenA were unable to assemble HIV contigs for these samples.

We have compared runtimes of all algorithms on the HIV-1 clinical data set in single-threaded and multi-threaded (32

threads) regimes, reporting average computation time (in seconds) to process one sample (see Supplementary Materials, Section 2.4).

HCV data set

First, we made a blind comparison of VirGenA, Vicuna and IVA accuracies on the HCV data set in the same way, as we had done for the HIV-1 clinical data set using our HCV vocabulary. Owing to high contamination by non-HCV reads in most of the samples, IVA was unable to assemble HCV contigs for 24 samples from O-Capt data set (Supplementary Table S11) and 37 samples from O-Meta data set (Supplementary Table S12), so we excluded IVA from comparisons on HCV data sets. Generally, VirGenA was concordant with Vicuna on HCV data set (Supplementary Tables S13 and S14); for O-Capt data set, the median cross-identity was 98.6%, and results for O-Meta data set are presented in Supplementary Materials Section 2.2 (Supplementary Tables S12, S14 and S15 and Supplementary Figure S5–S8). Our approach assembled longer contigs than Vicuna in comparison of all contigs assigned to a common reference and in comparison of contigs ascribed to references with maximal total coverage by assemblies (Supplementary Figures S9 and S10). Coverages of references by contigs were better for VirGenA (Supplementary Figures S11 and S12). VirGenA detected a single declared genotype in 14 of 22 ‘single-component’ HCV samples. The other eight samples were two subtype mixtures where the major component was declared and the frequency of the minor component was in the range of 0.05–17% (Supplementary Table S16). For all of these samples, assemblies of Vicuna also included contigs assigned to the second minor genotype detected by VirGenA (Supplementary Table S11). We then compared assemblies with the gold standard reference set: for 20 of 22 samples, both Vicuna and VirGenA covered gold standard references >99% of their length with identities

Figure 2. Comparison of assemblers on clinical dataset. For each sample a minimal reference set was constructed; all contigs were assigned to respective references from the minimal set. The same references for different samples were considered separately. (a) Lengths of longest contigs among all contigs assigned to each reference. Y-axis shows a total number of references in union of minimal reference sets of all samples. (b) Coverages of references each of which had maximal average coverage by contigs assembled by all methods for each sample in the dataset. (c) Average identities between all pairs of assemblies of each sample produced by different methods. (d) Enlarged scale for high values of the diagram shown on the panel (c). A colour version of this figure is available at BIB online: <https://academic.oup.com/bib>.

Figure 3. Comparison of assemblers on artificial two-component mixtures of divergent HIV strains. Lengths of longest contigs among all contigs assigned to each reference for each sample from data sets: PIRS_min (A), PIRS_median (C) and MIX (E). Identities of assemblies of components to corresponding references for each sample from data sets: PIRS_min (B), PIRS_median (D) and MIX (F). On panels A, C and E, Y-axes show total numbers of components in all samples together. On panels B, D and F, Y-axes show total numbers of components for all samples together normalized to one, which have nonzero coverage by contigs assembled by each software. A colour version of this figure is available at BIB online: <https://academic.oup.com/bib>.

>99%; the other two samples had lower coverage of their gold standard references by contigs, but identities of assemblies were also >99%. In two samples (sP883026 and sP731482), the second component had identities >99% with gold standard reference sequences of one sample (sP681788) of high viral load, and thus, likely were cross-contaminations (Supplementary Table S17).

Performance of assemblers on HIV artificial mixtures

To assess performance of each algorithm on mixtures of highly divergent variants, we prepared two-component mixtures with various proportions of components. Three data sets were prepared (see Supplementary Materials, 1.1, for details): samples in two data sets (further referred to as PIRS_median and PIRS_min) consisted of simulated samples with reads generated by pIRS [39] for three possible two-component mixtures of variants (HIV-1 subtypes A1, B and C). None of these variants matched any sequence in our HIV vocabulary reference set. PIRS_median contains references having median intrasubtype identity values to VirGenA vocabulary reference set, while PIRS_min comprises sequences least similar to sequences in VirGenA vocabulary reference set. Reads for the samples composing the third data set (further referred to as MIX) were proportionally picked from two of three selected clinical samples.

All assemblers were evaluated on these data sets. The same analyses as we had used for HIV clinical data set were done for

comparison of four approaches on simulated mixtures. For each assembly, we ascribed each contig to a reference from a set of two known virtual components of a mixture (see ‘Methods’ section).

Three approaches of read assembly were assessed by means of comparing lengths of longest contigs ascribed to one of two references, reference coverages by contig alignments, identities of contigs to ascribed references and identities to another reference (see Supplementary Materials, 1.8, for details, Supplementary Tables S18–S20 for PIRS_min data set, Supplementary Tables S21–S23 and Supplementary Tables S24–S26 for MIX data set). For simulated mixtures, VirGenA selected two references from the vocabulary and correctly assigned subtypes to major (or both) mixture components in most cases (Table 1), having best performance on MIX data set and worst on PIRS_min. Here, we emphasize that VirGenA did not know the correct answer when it was assembling reads because vocabulary HIV reference set did not include sequences corresponding to mixture components.

Frequencies of mixture components estimated by VirGenA were close to actual frequencies (Supplementary Tables S27–S29).

VirGenA outperformed other approaches in comparison of lengths of longest contigs that were ascribed to mixture components of simulated samples (Figure 3A, C and E). VirGenA assemblies for most samples contained near-complete-genome size ($8000 < \text{length} < 10000$) contigs for both components of

Table 1. Genotyping of artificial two-component mixtures of diverse HIV-1 strains by VirGenA

Data set	Number of samples	Number (%) of identified two-component samples	Number (%) of samples with correctly identified subtype of a major component	Number (%) of samples with both correctly identified subtypes of components
MIX	27	27 (100)	27 (100)	27 (100)
PIRS_median	108	108 (100)	106 (98)	101 (94)
PIRS_min	108	99 (92)	106 (98)	98 (91)

Table 2. Lengths of longest contigs assembled for components of artificial two-component mixtures of HIV-1 divergent strains

Data set	Number of samples	Number (%) of samples where lengths of longest contigs for both components are within (8000 or 10 000)		
		VirGenA	Vicuna	IVA
MIX	27	16 (59)	0 (0)	0 (0)
PIRS_median	108	69 (64)	1 (1)	14 (13)
PIRS_min	108	62 (57)	0 (0)	49 (45)

mixtures, whereas other assemblers tend to produce shorter (incomplete) or longer contigs (chimeric, containing fragments of different components) in assemblies (Table 2).

We compared assemblers as to how well alignments of their contigs covered two virtual components of simulated mixtures (Supplementary Figures S13–S15). VirGenA has slightly higher medians of coverage distributions across components of all samples than competitors (Table 3), but assembled contigs for mixtures sufficiently more accurately than other programs: significantly more mixture components collected from all samples were covered by contigs with identities of local alignments >99% on average produced by VirGenA compared with IVA and Vicuna (Figure 3B, D and F; Table 4). Although Vicuna had coverage statistics comparable with VirGenA, it was inferior in identity (and so did IVA), allowing us to suggest that it was because Vicuna and IVA produced more chimera contigs compared with VirGenA.

We estimated performances of three assemblers on a simulated three-component mixture that was generated by picking 50 000 read pairs in equal shares from the same three samples that were used for generating samples in MIX data set. VirGenA identified the mixture as containing three components and for each component selected a reference of correct subtype. Comparison between assemblers was organized in the same way as for two-component mixtures (Supplementary Table S30). VirGenA again outperformed competitors by all criteria.

Performance of assemblers on mixtures of HCV genotypes

We compared VirGenA subtypes detected for O-Capt data set of HCV-mixtures with those reported by Thompson et al. [30] (Supplementary Table S31). We successfully detected subtypes of both components for five of nine (IVT1–IVT9) controlled artificial mixtures where proportions of minor to major components were in range 1:1–1:50; for four other samples, only subtypes of major components were detected where proportions of components were 1:500 and 1:5000, and Vicuna assembled contigs that were assigned to both declared subtypes for eight of nine controlled mixtures, but in all cases, when VirGenA failed to detect one or both declared subtypes, Vicuna assembled only short (length < 300 bp) contigs for corresponded

Table 3. Median coverage of components by contigs of artificial two-component mixtures of divergent HIV-1 strains

Data set	VirGenA (%)	Vicuna (%)	IVA (%)
MIX	99.9	97.8	84.0
PIRS_median	100.0	98.0	99.9
PIRS_min	100.0	98.9	99.9

subtypes. The presence of only such short fragments was insufficient for genotyping by VirGenA (Supplementary Table S11).

We also detected both declared components in four of five clinical mixture samples (sS786324, sS589393 and JW1–JW3) (Supplementary Table S31). In the sample sS589393, only the major component was detected. Vicuna assembled contigs that were assigned to the same subtypes as detected by VirGenA for all but one JW3 sample for which both subtypes were represented by contigs slightly >300 bp. Two samples (HCVG10-02 and HCVG10-04) were not genotyped by VirGenA. Both programs assembled only short contigs (<300 bp) for these samples (Supplementary Table S11).

Estimated proportions of components for 8 of 16 HCV mixtures corresponded well with proportions of declared components (Pearson correlation > 0.99) (Supplementary Table S31). For O-Meta data set, results are presented in Supplementary Materials Section 2.3 (Supplementary Tables S12 and S32).

Assembly of recombinant HIV sequences

We estimated how VirGenA assembled a recombinant genotype using reference sequences of parental subtypes only. We generated artificial reads by pIRS for 265 sequences of HIV recombinant forms from HIV-1 reference set 2009 (Supplementary Table S33); for each recombinant, the set of parental subtypes was taken from <https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html>. We assembled these sequences *de novo* by IVA, Vicuna and by VirGenA using as a vocabulary NCBI Viral Genotyping Tool HIV12005 ‘pure’ reference set (<https://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi>), which did not contain recombinants. VirGenA ascribed unknown recombinant sequences to one to three ‘pure’ reference sequences, which usually were of parental subtypes; only for three samples, the sets of VirGenA subtypes were not strict subsets of

Table 4. Identity of local alignments of contigs and artificial components of mixtures of divergent HIV-1 strains

Data set	Total number of components	Number of covered components			Number (%) of components covered by contigs with average identities of local alignments >99%		
		VirGenA	Vicuna	IVA	VirGenA	Vicuna	IVA
MIX	54	54	54	54	51 (94)	25 (46)	37 (69)
PIRS_median	216	216	216	169	202 (94)	163 (75)	131 (78)
PIRS_min	216	211	216	192	211 (100)	168 (78)	190 (99)

corresponding sets of parental subtypes (Supplementary Table S34). About 15% of the samples were classified as mixtures; if an assembly was done following multiple references, it consisted of several contigs. For almost all samples, the longest contigs assembled by VirGenA were >8000 bp; thus, for each sample, a single reference was used for assembly of the major part of the genome (Supplementary Table S35, Supplementary Figure S16). It was surprising that VirGenA outperformed Vicuna in comparison of longest contigs; median lengths were 8500 and 7600 correspondingly. All three assemblers produced contigs, which covered genomes by >96% with identities >99% (Supplementary Table S35).

Discussion

Highly variable viruses establish divergent populations within hosts; also, mixed infections by multiple subtypes of viruses are not rare, especially in risk groups. The traditional approach to assembling consensus sequence by mapping individual reads onto reference sequence could potentially produce misalignment errors in divergent regions of a genome, which could be called and interpreted as genetic variants. We developed VirGenA—viral genome reference-guided assembler—that maps reads on multiple alignment of reference sequences, and then assembles them into contigs that are used for selecting a minimal set of optimal references for accurate reference-guided assembly.

We performed a comparison of our approach and two *de novo* assemblers, Vicuna [26] and IVA [27], specially designed for highly variable viral genomes such as HIV, HCV and HBV. Comparing assemblies of highly variable viral genomes from short reads is not a trivial task if the ‘true’ answer is not known. Here, we present a methodology that enables us to compare different assemblies based on several descriptive statistics. We believe that our methodology of comparison of short-read assemblies could be generally used.

The main advantage of our approach is the ability to account for the haplotype information obtained from the set of reference sequences, which represent the main genetic groups on viral phylogeny. Indeed, variants (SNVs) that are not covered by the same reads could be hardly phased without a priori knowledge. The phylogenetic tree often defines strong haplotypes between some SNVs, which jointly differentiate genetic groups. Thus, we may expect these haplotypes to present in viral quasispecies. Any assembly method, which is based solely on local information, more likely leads to chimera contigs, short contigs and/or low coverage of minor variants [43].

Our application is not suitable for reconstruction of similar haplotypes (quasispecies); it will restore their mixture in one consensus, ascribing then to one subtype. There are several

applications [48–50] that are specially designed for such problems. Our application may be useful for preparing or initial read alignments for further analysis by these specialized applications.

The reference-based assembling is potentially prone to errors in the cases of weak references. Practically, it means one of two situations: a far-away reference or unknown recombinant is sequenced. For the first, we simulated two-component mixtures of HIV subtypes least similar to references in the vocabulary (PIRS_min, see section ‘Performance of assemblers on HIV artificial mixtures’), and the median distances of artificial components to the most similar references from the vocabulary were ~11%, which were much higher than mean intrasubtype divergences for HIV (see Supplementary Materials, Section 1.1). For the second, we simulated NGS reads for HIV circulated recombinant forms and assembled them using a set of ‘pure’ subtypes (see section ‘Assembly of recombinant HIV sequences’). For both complicated for our approach cases, we obtained reasonably good results.

For samples that are mixtures of genotypes whose divergences are higher than the specified threshold, our approach produces correct results of a sample genotyping if a vocabulary used for assembly includes representing sequences (Table 1, Supplementary Table S27). The genetic diversity for most clinically important viruses like HIV, HCV and HBV is well characterized including circulating recombinant forms (<https://www.hiv.lanl.gov/>) [44, 45], so preparing reference vocabularies is a technical task. If a recombinant form is not represented in the VirGenA vocabulary, then assembly may be fragmented according to parental genotypes and interpreted as a mixture, however assembled contigs cover the sequenced recombinant genome well with high identities. Moreover, in most cases VirGenA assembles complete or near complete-genome contig following a single selected reference of one of parental subtypes (Supplementary Table S35, Supplementary Figure S16).

We use the chimera correction procedure that prevents forming of artificial chimera contigs in cases of mixtures. It cuts out long identical subsequences from contigs having of lower reads’ coverage and ascribes of corresponding reads to the contig with the highest coverage. The procedure also does not produce errors in the case of assembling of an unknown recombinant. All short contigs may be scaffolded, e.g. by SSPASE [38] (see Supplementary Methods, 1.7), to establish their correct linkage on a base of paired read information as well as contigs, which are assembled by *de novo* approaches like IVA and Vicuna.

It is interesting that assemblers demonstrated different performances for different viruses: on the HIV data set, IVA and VirGenA clearly outperformed Vicuna, but on the HCV data set, IVA with default settings failed to assemble HCV contigs for

multiple samples, on which Vicuna and VirGenA performed well. VirGenA correctly disentangles mixtures of strains belonging to divergent subtypes even in cases of extremely low ~1% admixtures and equal portions of components. We unexpectedly observed that about 12% of samples in the HIV-1 data set sequenced in the Sanger Institute, some of which previously had been used for comparison of *de novo* assemblers IVA and Vicuna [27], were mixtures of two or three subtypes. We cannot give exhaustive interpretation of these findings because no additional information is available for these samples, but we suspect that in most cases, admixtures are results of cross-sample contaminations, as in most mixed samples, low frequencies of minor variants (<1%) are observed; in 42% of mixtures, assembled sequences of minor variants have high alignment identity (>99%) with assemblies of major variants in other samples or single-component samples, and most of these samples (75%) are likely to be obtained from the same sequencing run, as samples to which these admixtures are similar. We also found clear evidence of contamination of two samples in the HCV data set by the third sample having high viral load. The scale of the problem of cross-sample contamination for the next-generation sequencing is not studied well, and probably the issue is likely to be overlooked; consequently, results of some studies that used ultra-deep sequencing for analyzing minor allele frequencies may be compromised.

Key Points

- VirGenA is a new viral reference-guided assembler, which can separate and genotype mixtures of strains of different intraspecies genetic groups (genotypes, subtypes, clades, etc.) and assemble a separate consensus sequence for each group in a mixture.
- VirGenA produces long assemblies for mixture components of extremely low frequencies (<1%) allowing detection of cross-contamination of samples by divergent genotypes.
- We observed that some clinical HIV and HCV samples in SRA are mixtures of two or three subtypes and provided evidence of cross-sample contaminations. VirGenA can be used to detect such contaminations.

Acknowledgements

Authors are grateful to Dr Bazykin G.A, Flyagin K.V., Patricia Palmer and Popova A.V. for useful discussions and help with the manuscript.

Funding

The Russian Industry and Trade Ministry (grant number 14411.2049999.19.015) and the National Institutes of Health, USA (grant numbers P30 CA006973 and UH2 CA211396-01).

Supplementary Data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

References

1. Department of Health and Human Services. Panel on anti-retroviral guidelines for adults and adolescents. *Guidelines for the Use of Antiretroviral Agents in HIV-1-infected Adults and Adolescents*. Department of Health and Human Services. <http://www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf>. 2016.
2. Terrault NA, Bzowej NH, Chang K-M, et al. AASLD guidelines for treatment of chronic hepatitis B. *Hepatology* 2016; **63**:261–83.
3. AASLD/IDSA HCV Guidance Panel. Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus: HEPATITIS C VIRUS GUIDANCE PANEL. *Hepatology* 2015; **62**:932–54.
4. Li JZ, Kuritzkes DR. Clinical implications of HIV-1 minority variants. *Clin Infect Dis* 2013; **56**:1667–74.
5. Li JZ, Paredes R, Ribaud HJ, et al. Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS* 2012; **26**:185–92.
6. Li JZ. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* 2011; **305**:1327.
7. Nicot F, Sauné K, Raymond S, et al. Minority resistant HIV-1 variants and the response to first-line NNRTI therapy. *J Clin Virol* 2015; **62**:20–4.
8. Gianella S, Delport W, Pacold ME, et al. Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. *J Virol* 2011; **85**:8359–67.
9. Cibulskis K, McKenna A, Fennell T, et al. ContEst: estimating cross-contamination of human samples in next generation sequencing data. *Bioinformatics* 2011; **27**:2601–2.
10. Bergmann EA, Chen B-J, Arora K, et al. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* 2016; **32**:3196–8.
11. Zanini F, Brodin J, Albert J, et al. Error rates, PCR recombination, and sampling depth in HIV-1 whole genome deep sequencing. *Virus Res* 2016, <https://doi.org/10.1016/j.virusres.2016.12.009>.
12. Wagner GA, Pacold ME, Kosakovsky Pond SL, et al. Incidence and prevalence of intrasubtype HIV-1 dual infection in at-risk men in the United States. *J Infect Dis* 2014; **209**:1032–8.
13. Redd AD, Quinn TC, Tobian AA. Frequency and implications of HIV superinfection. *Lancet Infect Dis* 2013; **13**:622–8.
14. Cappy P, De Oliveira F, Gueudin M, et al. A multiplex PCR approach for detecting dual infections and recombinants involving major HIV variants. *J Clin Microbiol* 2016; **54**:1282–8.
15. van Asten L, Prins M. Infection with concurrent multiple hepatitis C virus genotypes is associated with faster HIV disease progression. *AIDS* 2004; **18**:2319–24.
16. Pham ST, Bull RA, Bennett JM, et al. Frequent multiple hepatitis C virus infections among injection drug users in a prison setting. *Hepatology* 2010; **52**:1564–72.
17. Beggel B, Neumann-Fraune M, Doring M, et al. Genotyping hepatitis B virus dual infections using population-based sequence data. *J Gen Virol* 2012; **93**:1899–907.
18. van der Kuyl AC, Zorgdrager F, Hogema B, et al. High prevalence of hepatitis B virus dual infection with genotypes A and G in HIV-1 infected men in Amsterdam, the Netherlands, during 2000–2011. *BMC Infect Dis* 2013; **13**.
19. EASL. EASL recommendations on treatment of hepatitis C 2016. *J Hepatol* 2017; **66**:153–94.

20. Archer J, Rambaut A, Taillon BE, et al. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time—an ultra-deep approach. *PLoS Comput Biol* 2010;**6**:e1001022.
21. Pacold M, Smith D, Little S, et al. Comparison of methods to detect HIV dual infection. *AIDS Res Hum Retroviruses* 2010;**26**:1291–8.
22. Otto TD, Sanders M, Berriman M, et al. Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* 2010;**26**:1704–7.
23. Iqbal Z, Caccamo M, Turner I, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* 2012;**44**:226–32.
24. Henn MR, Boutwell CL, Charlebois P, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012;**8**:e1002529.
25. Willerth SM, Pedro HAM, Pachter L, et al. Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS One* 2010;**5**:e13564.
26. Yang X, Charlebois P, Gnerre S, et al. De novo assembly of highly diverse viral populations. *BMC Genomics* 2012;**13**:475.
27. Hunt M, Gall A, Ong SH, et al. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics* 2015;**31**:2374–6.
28. Wilson EM, Rosenthal ES, Kattakuzhy S, et al. Clinical laboratory testing in the era of directly acting antiviral therapies for hepatitis C. *Clin Microbiol Rev* 2017;**30**:23–42.
29. Lessells R, Katzenstein D, de Oliveira T. Are subtype differences important in HIV drug resistance? *Curr Opin Virol* 2012;**2**:636–43.
30. Thomson E, Ip CLC, Badhan A, et al. Comparison of next-generation sequencing technologies for comprehensive assessment of full-length hepatitis C viral genomes. *J Clin Microbiol* 2016;**54**:2470–84.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
32. Lee W-P, Stromberg MP, Ward A, et al. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One* 2014;**9**:e90581.
33. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.
34. Fantin YS, Neverov AD, Favorov AV, et al. Base-calling algorithm with vocabulary (BCV) method for analyzing population sequencing chromatograms. *PLoS One* 2013;**8**:e54835.
35. Rozanov M, Plikat U, Chappey C, et al. A web-based genotyping resource for viral sequences. *Nucleic Acids Res* 2004;**32**:W654–9.
36. Parrish N, Hormozdiari F, Eskin E. Assembly of non-unique insertion content using next-generation sequencing. *BMC Bioinformatics* 2011;**12**:S3.
37. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
38. Boetzer M, Henkel CV, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 2011;**27**:578–9.
39. Hu X, Yuan J, Shi Y, et al. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* 2012;**28**:1533–5.
40. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
41. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
42. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
43. Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, et al. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 2014;**15**:37.
44. Shin-I T, Tanaka Y, Tateno Y, et al. Development and public release of a comprehensive hepatitis virus database. *Hepatol Res* 2008;**38**:234–43.
45. Lau KA, Wong JYL. Current trends of HIV recombination worldwide. *Infect Dis Rep* 2013;**5**:e4.