Ji H et al. *Journal of the International AIDS Society* 2018, **21**:e25193
http://onlinelibrary.wiley.com/doi/10.1002/jia2.25193/full | https://doi.org/10.1002/jia2.25193

JIAS
JOURNAL OF THE INTERNATIONAL AIDS SOCIETY

COMMENTARY

Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: the Winnipeg Consensus

Hezhao Ji[1,2§], Eric Enns[3], Chanson J. Brumme[4], Neil Parkin[5], Mark Howison[6], Emma R. Lee[1], Rupert Capina[1], Eric Marinier[3], Santiago Avila-Rios[7], Paul Sandstrom[1,2], Gary Van Domselaar[2,3], Richard Harrigan[8], Roger Paredes[9], Rami Kantor[10] and Marc Noguera-Julian[9§]

§Corresponding authors: Hezhao Ji, National HIV & Retrovirology Laboratories, Public Health Agency of Canada, Winnipeg, Canada. Tel: +1 204 7896521. (hezhao.ji@canada.ca)
Marc Noguera-Julian, IrsiCaixa AIDS Research Institute, Badalona, Catalonia, Spain. Tel: + 34 934656374. (mnoguera@irsicaixa.es)

Abstract

Introduction: Next-generation sequencing (NGS) has several advantages over conventional Sanger sequencing for HIV drug resistance (HIVDR) genotyping, including detection and quantitation of low-abundance variants bearing drug resistance mutations (DRMs). However, the high HIV genomic diversity, unprecedented large volume of data, complexity of analysis and potential for error pose significant challenges for data processing. Several NGS analysis pipelines have been developed and used in HIVDR research; however, the absence of uniformity in data processing strategies results in lack of consistency and comparability of outputs from different pipelines. To fill this gap, an international symposium on bioinformatic strategies for NGS-based HIVDR testing was held in February 2018 in Winnipeg, Canada, convening laboratory scientists, bioinformaticians and clinicians involved in four recently developed, publicly available NGS HIVDR pipelines. The goal of this symposium was to establish a consensus on effective bioinformatic strategies for NGS data management and its use for HIVDR reporting.

Discussion: Essential functionalities of an NGS HIVDR pipeline were divided into five analytic blocks: (1) NGS read quality control (QC)/quality assurance (QA); (2) NGS read alignment and reference mapping; (3) HIV variant calling and variant QC; (4) NGS HIVDR reporting; and (5) extended data applications and additional considerations for data management. The consensuses reached among the participants on all major aspects of these blocks are summarized here. They encompass not only recommended data management and analysis strategies, but also detailed bioinformatic approaches that help ensure accuracy of the derived HIVDR analysis outputs for both research and potential clinical use.

Conclusions: While NGS is being adopted more broadly in HIVDR testing laboratories, data processing is often a bottleneck hindering its generalized application. The proposed standardization of NGS read QC/QA, read alignment and reference mapping, variant calling and QC, HIVDR reporting and relevant data management strategies in this "Winnipeg Consensus" may serve as a starting guideline for NGS HIVDR data processing that informs the refinement of existing pipelines and those yet to be developed. Moreover, the bioinformatic strategies presented here may apply more broadly to NGS data analysis of microbes harbouring significant genomic diversity.

Keywords: next-generation sequencing; HIV drug resistance test; bioinformatics; pipeline; Winnipeg Consensus; guideline

Received 18 May 2018; **Accepted** 26 September 2018

1 | INTRODUCTION

Successful antiretroviral therapy (ART) suppresses HIV viral load, reduces the incidence of new infections and increases the life expectancy of infected individuals [1-5]. However, HIV drug resistance (HIVDR) can occur as result solely from poor proofreading during viral replication or the combined effect from poor proof-reading and drug selection during unsuccessful ART [6,7]. With drastic increase in ART coverage worldwide, HIVDR has become a major barrier that hinders its effectiveness [8]. Conventional HIVDR genotyping qualitatively detects drug resistance mutation (DRM) using Sanger sequencing approaches, which has limited capacity in reliable detection of minority variants present at frequencies below approximately 20%, with potentially relevant clinical impact [9-11].

Next-generation sequencing (NGS), as exemplified by Illumina sequencing-by-synthesis technology, refers to newer sequencing technologies that enable high-throughput,

massively parallel sequencing of individual input templates [11-13]. When applied to HIVDR genotyping, such technologies bestow unique advantages and significantly improve sensitivity for resolving complex HIV quasispecies with exceptional resolution and quantitative minority variant identification [11,13,14]. The high scalability and ongoing cost reduction of NGS also permit further improvement in time efficiency and cost-effectiveness of NGS HIVDR assays when many batched specimens are being processed [15-17]. While broader adoption in testing laboratories could lead to new NGS-based standards for HIVDR genotyping, some important issues remain to be addressed, including lack of standardization for NGS HIVDR data analysis pipelines and resulting accurate and meaningful low-abundance variant data interpretation [11,13,18].

Like other molecular assays, the routine use of NGS HIVDR assays requires fully validated protocols that dictate sample processing in the laboratory. However, NGS also requires well-defined bioinformatics strategies and tools that help to reliably convert raw NGS data into user-interpretable HIVDR results. Notably, with the broad adoption of NGS, the sequencing itself has become relatively less challenging, while the data processing steps have become the primary bottleneck for its generalized application to HIVDR. Such challenges arise largely from: (1) high HIV sequence diversity [19]; (2) unprecedented large volume of NGS data, (3) sequence-specific errors, some of which are intrinsic to different NGS platforms [20,21]; (4) relatively short NGS read lengths with suboptimal basecalling accuracies; and (5) requirement for advanced bioinformatics skills and high performance computing capacity. Most NGS software applications are designed for the analysis of organismal genomes of a fixed ploidy and having modest sequence coverage. In contrast, the HIV genome exists as a quasispecies, and thus presents unique challenges for its sequencing and analysis. Existing NGS analysis pipelines for HIVDR to date have been developed by independent research groups with little coordination or any pre-existing guidelines to reference, and thus differ in their data processing strategies and their output formats (Table 1). This lack of conventions to which to adhere leads to uncertainties in data reliability and also makes the comparison of outputs from different pipelines unnecessarily difficult [11]. Moreover, it also impedes the ability of regulatory agencies to standardize and benchmark such assays for accreditation purposes. Thus, a consensus recommendation on standards for bioinformatic analysis and reporting conventions for HIVDR research and clinical purposes is urgently required.

Development of such a consensus necessitates knowledge of NGS data characteristics, relevant bioinformatics skill sets, appreciation of the clinical relevance (or lack thereof) of minority variants and, importantly, extensive expertise and experience in performing NGS HIVDR data analysis. In this commentary, we report the outcome of an international symposium on bioinformatic strategies for NGS HIVDR testing, which was held in February 2018 in Winnipeg, Canada, convening bioinformaticians, scientists and clinicians from four NGS HIVDR pipeline teams, including: HyDRA from the National Microbiology Laboratory in Canada, PASeq.org from Institute for AIDS Research (IrsiCaixa) in Spain, MiCall from the British Columbia Centre for Excellence in HIV/AIDS in Canada and hivmmer from the Providence-Boston Center for AIDS Research at Brown University in USA. Notably, HyDRA, PASeq.org and MiCall are freely available web interfaces and are used by many investigators worldwide, while hivmmer and several other pipelines are also freely available but still require advanced computational skills to execute (Table 1). In-depth discussions and brainstorming sessions were organized during the symposium. The consensus for NGS-based HIVDR data analysis that was reached among the participating groups (referred to as the "Winnipeg Consensus" hereafter) is summarized and presented here. It is noteworthy that all bioinformatics strategies discussed at the symposium and presented in this "Winnipeg Consensus" are based on the second-generation sequencing technologies exemplified by Illumina sequence-by-synthesis technology.

2 | DISCUSSION

The characteristics of an optimal NGS HIVDR data processing pipeline include: (1) automated data analysis with a short turnaround time; (2) accommodation of all relevant HIV genes and raw data from varied NGS platforms; (3) incorporation of essential quality assurance (QA)/quality control (QC) strategies to ensure data accuracy and reproducibility; (4) production of customizable and easy-to-interpret HIVDR reports that satisfy research, surveillance and clinical monitoring needs; (5) user-friendliness requiring minimal or no bioinformatics experience; and (6) easy access with minimal additional cost to the end-users. The Winnipeg Consensus covers the major bioinformatic strategies that help to satisfy these requirements.

Although pipelines vary, some basic principles apply in NGS HIVDR data analysis. The analytic components of an NGS HIVDR pipeline were grouped into five sequential functional blocks: (1) NGS read QC/QA; (2) NGS read alignment and reference mapping; (3) HIV variant calling and variant QC; (4) HIVDR interpretation and reporting; and (5) analysis data management. Table 2 details the Winnipeg Consensus on the major functionalities in each of these blocks, including analysis objectives, consensus on strategies and associated considerations, where applicable. The highlights include:

1 "**NGS read QC/QA**" warrants that only high-quality NGS reads are to be utilized in downstream HIVDR data analysis. Although all NGS platforms attach quality scores to individual basecalls, the additional NGS read QC/QA steps described in this consensus were deemed both necessary and effective in reducing false variant calling. Only basic read QC/QA strategies are described here and more stringent filtering may be required in certain cases.
2 "**NGS read alignment and reference mapping**" addresses the needs for valid and accurate read alignment to designated reference sequence(s) that enables subsequent variant calling. Pipelines should at minimum support reference mapping of the whole HIV *pol* gene, which encodes the three main drug-targeted HIV enzymes: protease (PR), reverse transcriptase (RT) and integrase (IN). Although not urgently required for HIVDR genotyping, it would be beneficial for pipelines to also accommodate *full-length* HIV reference alignment, since many users are adopting NGS for partial or full-length HIV sequencing beyond the *pol* gene. Notably, genetic variability in the HIV *env* gene poses more challenges for reference alignment strategies than the

Table 1. Currently available pipeline/software for automated NGS-based HIVDR data analysis

	Reference information		Resources			Technical characteristics			HIV drug resistance		HIVDR data analysis features		
Pipeline/software	URL	Year[c]	Cost[d]	Time[e]	Bioinformatic IT needs[f]	Compatible NGS platform	Cloud based	Web interface	Designed for HIVDR	Ref DB and algorithm[g]	Output (nt/*aa*)[h]	QA checks[i]	InDel[j]
V-Phaser 2 [40]	https://www.broadinstitute.org/viral-genomics/v-phaser-2	2013	Free	N/A	Yes	N/A	No	No	No	No	csv/*csv*	E	Yes
ShoRah [41]	https://github.com/cbg-ethz/shorah	2013	Free	N/A	Yes	N/A	No	No	No	No	csv/*N/A*	E	N/A
VirVarSeq [42]	https://sourceforge.net/projects/virtools/	2015	Free	N/A	Yes	Illumina	No	No	No	No	fasta/*csv*	Q/E	Yes
MinVar [43]	https://ozagordi.github.io/MinVar/	2016	Free	<1 hour	Yes	Illumina	Yes[k]	No	Yes	HIVdb	vcf/*csv*	Q	Yes
V-pipe	https://cbg-ethz.github.io/V-pipe/	2017	Free	N/A	Yes	Illumina	No	No	No	No	fasta/csv	Q	Yes
Hivmmer [44]	https://github.com/kantorlab/hivmmer	2017	Free	<1 hour	Yes	Illumina	No	No	Yes	No	csv/*csv*	L	Yes
Geno2Pheno[ngs-freq][45] [a]	http://ngs.geno2pheno.org/	2018	Free	<1 minute	Yes	N/A	No	Yes	Yes	g2p[res]	fasta/*csv*	N/A	Yes
MiCall[b]	https://github.com/cfe-lab/MiCall	2016	Free	<1 hour	No	Illumina	Yes	Yes	Yes	HIVdb	Csv/*csv*	Q/E	Yes
HyDRA	https://hydra.canada.ca	2016	Free	<1 hour	No	Illumina Ion Torrent	No	Yes	Yes	HIVdb	fasta,vcf/*aavf*	Q/E	Yes
PASeq.org	https://www.paseq.org	2016	Free	<1 hour	No	Illumina	Yes	Yes	Yes	HIVdb	fasta/*csv*	Q/C/A	Yes
DeepChek HIV [46]	https://www.ablsa.com/overview/deepchek/	2014	$65[l]	<1 hour	No	Illumina Ion Torrent	Yes	Yes	Yes	HIVdb	csv/*csv*	Q	Yes
Smart GeneHIV	http://www.smartgene.com/mod_ngs.html	2016	N/A	N/A	No	N/A	No	Yes	Yes	HIVdb	N/A	N/A	Yes
Vela sentosa HIV [47]	http://www.veladx.com/HIV.html	2016	$200[m]	N/A	No	Ion Torrent	No	Yes	Yes	HIVdb	fasta/*csv*	Q	Yes
Hyrax Exatype	https://exatype.com/	2018	N/A	<1 hour	No	Illumina IonTorrent	Yes	Yes	Yes	HIVdb	fasta/*csv*	Q	Yes

The pipelines/software are categorized as: (1) freely available software for bioinformaticians (top block); (2) freely available software suitable for non-bioinformaticians (middle block); and (3) commercial software (bottom block). Within each block, the chronological order was followed. N/A**,** Information not available or not applicable. NGS, next-generation sequencing; HIVDR, HIV drug resistance.

[a]Geno2pheno[ngs-freq] pipeline can only use a codon frequency table as an input which needs to be obtained separately; [b]pending approval and release on Illumina BaseSpace Sequence Hub. For early access, please micalldev@cfenet.ubc.ca; [c]year of publication/*public availability*; [d]approximate per sample cost of bioinformatic data analysis only; [e]time range for single sample data analysis (data transfer time excluded); [f]refers to the need of on-site computational infrastructure or expert staff; [g]Ref DB and algorithm: reference HIV resistance database and/or algorithms for HIV resistance interpretation *(HIVdb: Stanford HIV Database)*; g2p[res] refers to the Geno2pheno[resistance] statistical engine; [h]output: format of output files reporting nucleotide (nt) and amino acid (aa) variations; [i]QA check strategies incorporated for NGS read quality assurance (Q: Quality Control; C: Contamination Control; E: Sequencing Error Model; L: Alignment Quality Filter; A: ApoBEC Hypermutation Detection); [j]indels are recognizable by default but no codon-aware strategies are implemented for reporting insertion/deletion mutations specifically associated to HIV resistance; [k]can be ported to Cloud; [l]cost based on general access through Illumina basespace; [m]approximate cost of whole sample analysis (sample preparation, sequencing, data analysis).

relatively conserved *pol* gene. Certain insertions and deletions (indels) in HIV-1 PR (near codon 35) and RT (near codon 69) genes are associated with drug resistance and such indels should be identified and reported for both HIVDR surveillance and clinical monitoring purposes [22-25]. Identification of such indels at the final HIVDR reporting stage is a relevant outcome of this alignment and reference mapping step. Indel management strategies differ among existing pipelines (Table 1). While several pipelines claim to accommodate indels in variant calling and DRM detection, pipelines that use NGS short-read aligners such as bowtie2 [26] may not adequately address such needs, since short-read aligners cannot straightforwardly be used to capture the effect of indels on the resulting coding sequence. Other approaches that perform haplotype phasing or that incorporate codon-aware alignment strategies may be needed to reliably detect known HIVDR-associated indels, but further evaluation is needed.

3 "**HIV variant calling and variant QC**" imposes additional stringency on the calling of variants, which is especially important when minority variants are concerned. NGS errors may arise at multiple points during sample processing (e.g. nucleic acid extraction, reverse transcription, PCR, template amplicon preparation for NGS and NGS sequencing) and NGS data processing [27]. The gross error rates generated from short-read NGS platforms ranges from approximately 1 to 10 errors per 1000 bases leading to increased false positive detection of minority variants when their prevalence falls below approximately 1% [13,28-30]. The additional variant QC strategies significantly improve the reliability of calling variants of low abundance, undetectable by Sanger sequencing. It is acknowledged that the threshold of minority variant frequency considered to be clinically relevant remains debatable [31].

4 "**NGS HIVDR interpretation and reporting**" is the only component designed specifically for HIVDR application, while all other blocks and associated strategies may find broader application, especially for genomic sequence analysis of microbes harbouring high genomic diversity, similar to HIV. This specific element of the pipeline streamlines the strategies to convert valid NGS-derived amino acid variant data into end-user-interpretable HIVDR results. Two HIVDR report formats are recommended in this Consensus for addressing needs of either research-oriented projects (a *comprehensive report*) or clinically oriented testing (a *concise report*). Ultimately, a customizable HIVDR reporting strategy is preferred for an optimal pipeline, allowing the users to construct a report of their preference. To facilitate comparisons and merging of data from different pipelines, a new standard amino acid variant file (aavf) format has been proposed (Appendix 1, https://github.com/winhiv/aavf-spec). Based on the variant call format (vcf) standard that has been universally adopted for recording nucleotide variants, the aavf report provides a compact summary of the amino acid variation obtained by conceptual translation of the NGS read pileup across the examined region of the HIV genome. It also contains information on the frequencies of matching codons (wild type or mutant), quality of the variant calling as well as the coverage of relevant loci. Although the specification is designed to fully accommodate the requirements for reporting of NGS-based HIVDR testing, it is still suitably generic to serve as a general purpose file format for reporting amino acid variants for broader applications. A tool suite to parse aavf format is available at https://github.com/winhiv/PyAAVF.

5 "**General analysis data management**" deals with issues that concern both the data generator and the analysis provider, to protect the best interests of both parties, including formats and contents for data storage, software versioning, information traceability and data ownership policies.

This symposium was held at a time when NGS for HIVDR genotyping is increasingly being adopted by many laboratories for research, surveillance and clinical monitoring purposes. Although the functionalities and assembly of bioinformatics strategies applied in different pipelines vary, they share a common objective. The Winnipeg Consensus addresses the urgent needs for and starts the process of standardization of NGS HIVDR data analysis pipelines. It is noteworthy that most of the bioinformatics strategies described in the Winnipeg Consensus have already been incorporated in three of the assessed pipelines, which explains the high concordance among these pipelines when the same data sets were analysed [32]. Although minor differences currently exist among PASeq, HyDRA and MiCall regarding the data processing procedures and reporting strategies, preliminary data suggests that these pipelines are largely interchangeable especially when only HIVDR mutations present at ≥5% are of interest [32].

An additional important outcome of this symposium was a consensus that a well-characterized NGS HIVDR "dry panel" should be constructed in support of both pipeline development and validation applications. Such a dry panel would consist of a variety of simulated data files as well as empirical data sets derived from plasmids, artificial plasmid mixtures and patient specimens. It should also cover all major HIV-1 subtypes and signature DRMs at a wide range of frequencies, allowing the flexibility for end-users to customize panels based on their needs. Such a comprehensive panel is currently under construction by the symposium participant teams and will become freely accessible to the public once established. In fact, a subset of the dry panel has already been used for a comparison of PASeq, HyDRA and MiCall [32].

Additional NGS HIVDR assay comparative assessment strategies, such as parallel testing of the same plasma specimens in different laboratories followed by analysis of the raw NGS data from each laboratory using all available pipelines, are also underway. This is in collaboration with the Virology Quality Assurance (VQA) programme supported by the Division of AIDS at the National Institutes of Health, USA, which provides quality assurance support for HIVDR laboratories worldwide [33].

It is acknowledged that some limitations exist in the Winnipeg Consensus, including: (1) it only addresses strategic issues concerning NGS data processing and subsequent report accuracy. Errors arising from pre-analytical procedures remain to be minimized through comprehensive protocol validations [34]; (2) strategies described here ensure the quality of minority variant detection and reporting based solely on the input NGS data, thus assuming that the applied NGS reads directly represent the intrahost viral quasispecies. Understandably, the sensitivity and accuracy of NGS in minority variant quantification are inherently dependent on the initial HIV RNA template input,

Table 2. Outcomes of the Winnipeg Consensus for recommended bioinformatic strategies for an NGS-based HIVDR data analysis pipeline

Functional blocks	Objectives	Consensus on strategies	Notes and comments
1. NGS read quality control/quality assurance	To ensure only quality reads are applied in downstream data processing	1. Average quality score (QS) of the read: 25	A QS at 25 corresponds to an estimated sequencing error rate of 0.3% [48]. When possible, direct QS examination for all individual bases and exclusion of those with scores <25 from subsequent analysis are recommended
		2. Minimum read length: 75 bases	This is based on Illumina 300-cycle paired-end sequencing and it may vary if another NGS platform or sequencing protocol is applied
		3. Contamination check: recommended	External non-viral contamination may be interfering with HIV NGS efficiency. HIV cross-sample contamination or "index hopping" implies errors in laboratory sample processing which may lead to erroneous minority variant detection (see strategies implemented in V-Pipe and ViCroSeq[https://github.com/mnoguera/ViCroSeq] tools) [49]
		4. APOBEC mutation check: recommended	Presence of APOBEC-edited DNA templates in the sequenced sample may result in the artefactual detection of minority variant related to APOBEC activity. Filtering this non-viable sequence content is beneficial especially when significant amounts of HIV proviral DNA may be present in the input specimen (i.e. PBMC, dried blood spots) [50,51]
2. NGS read alignment and reference mapping	To ensure the efficiency and accuracy of NGS read alignment to reference sequences	1. HIV-1 reference: HXB-2	For conserved regions such as HIV-1 *pol*, the choice of reference has minimal impact on subsequent alignment to a single reference. HXB-2 is a natural choice for the reference sequence since it provides the standard coordinate system for reporting DRMs. Iterative realignment to a sample-specific consensus may also reduce the importance of the initial choice of reference sequence. However, for variable regions such as *env*, a more comprehensive collection/database of reference sequences should be evaluated
		2. NGS read aligner: short-read aligner is recommended	Bowtie2 is thus far the most commonly used NGS short-read aligner due to its speed, availability, documentation, ease of installation and active maintenance [26]. An alternative to NGS short-read alignment is to conduct probabilistic multiple-sequence alignment with HMMER [52]. Other aligners and alignment strategies that have been previously evaluated by the group but are no longer in use include SMALT, BWA-MEM, BLAST [53], custom implementations of codon-aware Smith–Waterman alignment [54], MOSAIK [55], stampy [56] and SHRiMP2 [57]
		3. Analysis of whole *pol* gene: required	Coverage of the entire *pol* region is required to enable HIVDR analysis on all genes encoding the three ART-targeted enzymes (protease, reverse transcriptase and integrase)
		4. Indel management: required	Effective indel management strategy (i.e. codon-aware alignments) is not available in existing pipelines. However, with several indel variations contributing to HIVDR, full-codon indels should be properly identified and reported

Table 2. (*Continued*)

Functional blocks	Objectives	Consensus on strategies	Notes and comments
3. HIV variant calling and variant quality control	To ensure the accuracy of variant calling	1. QC/QA of nucleotide variant calling: recommended	Additional QA/QC procedures may be incorporated to further ensure the variant call accuracy. For instance, HyDRA calls variation only when minimum allele counts is ≥5, minimum QS of variant allele is ≥30 and read depth at the relevant variation site is ≥100
		2. Amino acid variation calling based on NGS reads, but NOT consensus sequence: required	Consensus sequence-based DRM analysis renders inevitable assumption while ≥2 mixed bases present in the codon, which diminishes NGS values in quantitative DRM detections
		3. Secondary QC for minority variant calling: optional	It helps to exclude erroneous variant calls via statistical estimation based on gross platform-specific error rate, as determined by parallel sequencing of pedigreed plasmid in the NGS run, which sums all potential errors from any involved assay procedures
4. HIVDR interpretation and reporting	HIVDR interpretation	Query reference database and algorithms: HIVdb (https://hivdb.stanford.edu)	Although minor discrepancies exist with other alike databases, Stanford database (HIVdb) is recommended for better general adoption
	Concise report *(for potential clinical use)*	The report should contain the following: 1 Patient and sample information if provided (optional) 2 Exportable/printable HIVDR report with DRMs and colour-coded resistance interpretations 3 Two-column summary on DRMs at reporting threshold of 5% and 15% respectively with no detailed frequencies 4 Pipeline and software version applied 5 Comment on the accreditation status of the assay for clinical use	Concise HIVDR reports from NGS data should simulate Sanger sequencing output for easier adoption and interpretation by clinicians, to be used for clinical care Integrase gene should be examined in addition to reverse transcriptase and protease and samples with no integrase data should be flagged The reporting thresholds are suggested to simulate sensitivity of SS in DRM detection (15%) and to exemplify a practical threshold for reporting authentic DRMs of potential clinical relevance with minimal interference from errors/bias (5%). Further refinement of these values may be required as relevant research advances
	Comprehensive report *(for research use)*	The report should contain the following: 1 All contents included in clinical reports. 2 Summary on filtering statistics, quality metrics and coverage plots 3 Quantitative report on all HIV-1 DRMs with exact frequencies 4 Consensus sequence with threshold of 15% for mixed base call	Comprehensive reports should contain all NGS-derived data that researchers may utilize for various application purposes Customizable HIVDR reporting is encouraged to enable users to construct report that best serves their needs. For instance, a customizable frequency threshold setting for DRM identification and reporting and user-definable threshold(s) for consensus sequence generation are recommended
	Other exportable data *(Optional)*	1 Consensus sequences with user-defined threshold: recommended 2 Variant reports on all nucleotide loci: recommended 3 Variant reports on all amino acid loci: recommended 4 Codon usage at all amino acid variations loci: recommended	Standard VCF/BCF format is recommended for nucleotide variant reports. To facilitate comparisons and merging of data from different pipelines, a new standard .aavf reporting format is proposed (Appendix 1, https://github.com/winhiv/aavf-spec). A tool suite to parse aavf format is available at https://github.com/winhiv/PyAAVF. The aavf file provides an amino acid variation summary, along with frequencies of relevant codons, across the examined region based on the associated NGS reads directly. It may serve as a generic variation report template from any NGS analysis

Table 2. (*Continued*)

Functional blocks	Objectives	Consensus on strategies	Notes and comments
5. General Analysis data management	Data storage	1 Raw NGS data: to be stored by data generator while data analysis providers may transiently store it for reviewing purpose	Automated versioning of all analysis results, reports and intermediate data files is required for retroactive data assessments when necessary
		2 Intermediate files (*e.g. SAM, BAM*): no need to store	
		3 Versioning data files for the applied pipeline: recommended	
	Data disposal	Analysis provider disposes data after a defined holding period	Deposition of data into public archives (e.g. NCBI Sequence Read Archive) requires informed consent from the data generator
	Data ownership	1 Policy varies among different institutes and countries	In general, data generators own all the data while unidentified data may be used by data analysis provider for evaluation or research purposes providing mutual agreement is in place
		2 Clear data ownership statement should be included in Terms of Service and Conditions	

DRM, drug resistance mutation; NGS, next-generation sequencing; PBMC, peripheral blood mononuclear cell.

which in turn is defined by specimen characteristics and assay designs such as viral load, specimen volume processed, fraction of extracted nucleic acids used for RT-PCR, efficiency of RNA to DNA conversion and evenness of PCR amplification for HIV templates present in the specimen. Related accuracy limitations might be partially addressed using more sophisticated experimental designs such as primerID which is likely beneficial for research purposes, but not yet proven to be necessary for routine clinical use and hence not dealt with in this consensus [13,35-38]; and (3) it was developed primarily based on processing of data from Illumina technology, which is currently the most widely used, but not the only platform for NGS HIVDR [39]. Therefore, while Winnipeg Consensus principles apply to other NGS platforms, their exact implementation into data analysis pipelines will need to consider the platform-specific characteristics and sequence error profiles for optimal results [20].

3 | CONCLUSIONS

In conclusion, we present here the Winnipeg Consensus on bioinformatic strategies for NGS HIVDR data processing. This consensus may serve as an initial baseline to standardize NGS data analysis with a specific focus on HIVDR genotyping, and inform the refinement of existing pipelines and those still in development. This initiative and its subsequent activities may help make such technologies routine for both research and clinical HIVDR monitoring purposes, and may serve as a useful starting point for further developing of NGS analysis pipelines with similar and alternative intended applications.

AUTHORS' AFFILIATIONS

[1]National HIV and Retrovirology Laboratories at JC Wilt Infectious Diseases Research Centre, Public Health Agency of Canada, Winnipeg, MB, Canada; [2]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, MB, Canada; [3]Bioinformatics Core at the National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada; [4]British Columbia Centre for Excellence, HIV/AIDS, Vancouver, BC, Canada; [5]Data First Consulting Inc., Belmont, CA, USA; [6]Watson Institute for International and Public Affairs, Brown University, Providence, RI, USA; [7]Centre for Research in Infectious Diseases, National Institute of Respiratory Diseases, Mexico City, Mexico; [8]Division of AIDS, Department of Medicine, University of British Columbia, Vancouver, BC, Canada; [9]IrsiCaixa AIDS Research Institute, Badalona, Catalonia, Spain; [10]Division of Infectious Diseases, Brown University Alpert Medical School, Providence, RI, USA

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHORS' CONTRIBUTIONS

HJ, RP, PS, RH, GVD, RK and MNJ conceived and initiated the project. HJ, MNJ, NP, CJB and RK drafted this manuscript. All authors participated in the Winnipeg symposium and contributed to the manuscript revisions. MNJ, EE, CJB, MH, ERL, RC and EM led the discussions on varied bioinformatics issues at the symposium and summarized the consensus on the corresponding topics that were presented here. All authors contributed significantly to this study and have reviewed and approved the final version.

AUTHOR INFORMATION

The authors include members from three WHO Global HIV Drug Resistance Network (HIVResNet) member labs including BC Centre for Excellence in HIV/AIDS (CJB and RH) and the National HIV and Retrovirology Laboratories (HJ, PS, ERL and RC) from Canada and the National Institute of Respiratory

Diseases (SAR) from Mexico. RP, PS, RH, RK and NP serves as members of or advisors to the WHO HIVResNet. EE, MNJ, RH and MH are primary developers/coordinators of HyDRA, PASeq, MiCall and hivmmer programmes.

ACKNOWLEDGEMENTS

We appreciate the great in-kind support from all participating institutes and the financial supports from funding agencies, including the Canadian Federal Initiative to Address HIV and AIDS (PS, HJ, RC, ERL), Genomic Research and Development Initiative in Canada (HJ, PS, GVD, EE, EM). The development of MiCall was funded in part through the Genome British Columbia, Genome Quebec, Genome Canada and CIHR Partnership in Genomics and Personalized Health (Large-Scale Applied Research Project HIV142). The development of hivmmer was supported in part by the Providence/Boston Center for AIDS Research (P30AI042853), R01 AI108441 and the Brown University DEANS Award. RP and MNJ were supported in part by an unrestricted grant from Fundació CatalunyaCaixa – La Pedrera.

REFERENCES

1. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med. 2011;365(6):493–505.
2. Samji H, Cescon A, Hogg RS, Modur SP, Althoff KN, Buchacz K, et al. Closing the gap: increases in life expectancy among treated HIV-positive individuals in the United States and Canada. PLoS ONE. 2013;8(12):e81355.
3. Montaner JS, Reiss P, Cooper D, Vella S, Harris M, Conway B, et al. A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients: the INCAS Trial. Italy, The Netherlands, Canada and Australia Study. JAMA. 1998;279(12):930–7.
4. Hogg RS, Heath KV, Yip B, Craib KJ, O'Shaughnessy MV, Schechter MT, et al. Improved survival among HIV-infected individuals following initiation of antiretroviral therapy. JAMA. 1998;279(6):450–4.
5. Piot P, Quinn TC. Response to the AIDS pandemic – a global health model. N Engl J Med. 2013;368(23):2210–8.
6. Clutter DS, Jordan MR, Bertagnolio S, Shafer RW. HIV-1 drug resistance and resistance testing. Infect Genet Evol. 2016;46:292–307.
7. Tang MW, Shafer RW. HIV-1 antiretroviral resistance: scientific principles and clinical applications. Drugs. 2012;72(9):e1–25.
8. World Health Organization, United States Centers for Disease Control and Prevention, The Global Fund to Fight AIDS TaM. HIV Drug Resistance Report 2017. 2017.
9. Gunthard HF, Wong JK, Ignacio CC, Havlir DV, Richman DD. Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. AIDS Res Hum Retroviruses. 1998;14(10):869–76.
10. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. PLoS Med. 2008;5(7):e158.
11. Brumme CJ, Poon AFY. Promises and pitfalls of Illumina sequencing for HIV resistance genotyping. Virus Res. 2017;239:97–105.
12. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55(4):641–58.
13. Casadella M, Paredes R. Deep sequencing for HIV-1 clinical management. Virus Res. 2017;239:69–81.
14. Ji H, Liang B, Li Y, Van Domselaar G, Graham M, Tyler S, et al. Low abundance drug resistance variants in transmitted HIV drug resistance surveillance specimens identified using tagged pooled pyrosequencing. J Virol Methods. 2013;187(2):314–20.
15. Lapointe HR, Dong W, Lee GQ, Bangsberg DR, Martin JN, Mocello AR, et al. HIV drug resistance testing by high-multiplex "wide" sequencing on the MiSeq instrument. Antimicrob Agents Chemother. 2015;59(11):6824–33.
16. Ji H, Li Y, Graham M, Liang BB, Pilon R, Tyson S, et al. Next-generation sequencing of dried blood spot specimens: a novel approach to HIV drug-resistance surveillance. Antivir Ther. 2011;16(6):871–8.
17. Inzaule SC, Ondoa P, Peter T, Mugyenyi PN, Stevens WS, de Wit TFR, et al. Affordable HIV drug-resistance testing for monitoring of antiretroviral therapy in sub-Saharan Africa. Lancet Infect Dis. 2016;16(11):e267–75.
18. Chimukangara B, Samuel R, Naidoo K, de Oliveria, T. Primary HIV-1 drug resistant minority variants. AIDS Rev. 2017;19(2):89–96.
19. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science. 1995;267(5197):483–9.
20. Shin S, Park J. Characterization of sequence-specific errors in various next-generation sequencing systems. Mol BioSyst. 2016;12(3):914–22.
21. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.
22. Kozisek M, Saskova KG, Rezacova P, Brynda J, van Maarseveen NM, De Jong D, et al. Ninety-nine is not enough: molecular characterization of inhibitor-resistant human immunodeficiency virus type 1 protease mutants with insertions in the flap region. J Virol. 2008;82(12):5869–78.
23. Melikian GL, Rhee SY, Taylor J, Fessel WJ, Kaufman D, Towner W, et al. Standardized comparison of the relative impacts of HIV-1 reverse transcriptase (RT) mutations on nucleoside RT inhibitor susceptibility. Antimicrob Agents Chemother. 2012;56(5):2305–13.
24. Tramuto F, Bonura F, Mancuso S, Romano N, Vitale F. Detection of a new 3-base pair insertion mutation in the protease gene of human immunodeficiency virus type 1 during highly active antiretroviral therapy (HAART). AIDS Res Hum Retroviruses. 2005;21(5):420–3.
25. White KL, Chen JM, Margot NA, Wrin T, Petropoulos CJ, Naeger LK, et al. Molecular mechanisms of tenofovir resistance conferred by human immunodeficiency virus type 1 reverse transcriptase containing a diserine insertion after residue 69 and multiple thymidine analog-associated mutations. Antimicrob Agents Chemother. 2004;48(3):992–1003.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.
27. Beerenwinkel N, Gunthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Front Microbiol. 2012;3:329.
28. Pessoa R, Loureiro P, Esther LM, Carneiro-Proietti AB, Sabino EC, Busch MP, et al. Ultra-deep sequencing of HIV-1 near Full-length and partial proviral genomes reveals high genetic diversity among Brazilian blood donors. PLoS ONE. 2016;11(3):e0152499.
29. Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y, et al. Quasispecies analyses of the HIV-1 near-full-length genome with Illumina MiSeq. Front Microbiol. 2015;6:1258.
30. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of next generation sequencing platforms. Next Gener Seq Appl. 2014;1:1000106.
31. Stella-Ascariz N, Arribas JR, Paredes R, Li JZ. The role of HIV-1 drug-resistant minority variants in treatment failure. J Infect Dis 2017; 216 Suppl_9: S847–50.
32. Noguera-Julian M, Edgil D, Harrigan PR, Sandstrom P, Godfrey C, Paredes R. Next-generation human immunodeficiency virus sequencing for patient management and drug resistance surveillance. J Infect Dis 2017; 216 Suppl_9: S829–33.
33. Parkin N, Bremer J, Bertagnolio S. Genotyping external quality assurance in the World Health Organization HIV drug resistance laboratory network during 2007-2010. Clin Infect Dis. 2012;54 Suppl 4:S266–72.
34. McElroy K, Thomas T, Luciani F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. Microb Inform Exp. 2014;4(1):1.
35. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci USA. 2011;108(50):20166–71.
36. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43(6):e37.
37. Keys JR, Zhou S, Anderson JA, Eron JJ Jr, Rackoff LA, Jabara C, et al. Primer ID informs next-generation sequencing platforms and reveals preexisting drug resistance mutations in the HIV-1 reverse transcriptase coding domain. AIDS Res Hum Retroviruses. 2015;31(6):658–68.
38. Seifert D, Di Giallonardo F, Topfer A, Singer J, Schmutz S, Gunthard HF, et al. A comprehensive analysis of primer IDs to study heterogeneous HIV-1 populations. J Mol Biol. 2016;428(1):238–50.
39. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13:341.
40. Yang X, Charlebois P, Macalalad A, Henn MR, Zody MC. V-Phaser 2: variant inference for viral populations. BMC Genomics. 2013;14:674.
41. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics. 2011;12:119.
42. Verbist BM, Thys K, Reumers J, Wetzels Y, Van der Borght K, Talloen W, et al. VirVarSeq: a low-frequency virus variant detection pipeline for Illumina sequencing using adaptive base-calling accuracy filtering. Bioinformatics. 2015;31(1):94–101.

43. Huber M, Metzner KJ, Geissberger FD, Shah C, Leemann C, Klimkait T, et al. MinVar: a rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. J Virol Methods. 2017;240:7–13.
44. Howison M, Coetzer M, Kantor R. Measurement error and variant-calling in deep Illumina sequencing of HIV. bioRxiv 2018. https://doi.org/10.1101/276576
45. Döring M, Büch J, Friedrich G, Pironti A, Kalaghatgi P, Knops E, et al. geno2-pheno[ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. Nucleic Acids Res. 2018;46:W271–7.
46. Garcia-Diaz A, McCormick A, Booth C, Gonzalez D, Sayada C, Haque T, et al. Analysis of transmitted HIV-1 drug resistance using 454 ultra-deep-sequencing and the DeepChek((R))-HIV system. J Int AIDS Soc. 2014;17 4 Suppl 3:19752.
47. Chan M, Smirnov A, Mulawadi F, Lim P, Lim W-H, Leong ST, et al. A novel system control for quality control of diagnostic tests based on next-generation sequencing. J Appl Lab Med. 2016;1(1):25–35.
48. Illumina. Understanding Illumina Quality Scores. 2014.
49. Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis. 2017.
50. Dauwe K, Staelens D, Vancoillie L, Mortier V, Verhofstede C. Deep sequencing of HIV-1 RNA and DNA in newly diagnosed patients with baseline drug resistance showed no indications for hidden resistance and is biased by strong interference of hypermutation. J Clin Microbiol. 2016;54(6):1605–15.
51. Noguera-Julian M, Cozzi-Lepri A, Di Giallonardo F, Schuurman R, Daumer M, Aitken S, et al. Contribution of APOBEC3G/F activity to the development of low-abundance drug-resistant human immunodeficiency virus type 1 variants. Clin Microbiol Infect. 2016; 22(2): 191–200.
52. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7 (10):e1002195.
53. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
54. Gianella S, Delport W, Pacold ME, Young JA, Choi JY, Little SJ, 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(16):8359–67.
55. Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS ONE. 2014;9(3):e90581.
56. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936–9.
57. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: sensitive yet practical short read mapping. Bioinformatics. 2011;27(7):1011–2.

APPENDIX 1

1. The AAVF specification

AAVF is a text file format, inspired by the Variant Call Format (VCF) format. It contains meta-information lines, a header line, and then data lines each containing information about a position in a gene within a genome.

1.1 An example

```
##fileformat=AAVFv1.0
##fileDate=20180501
##source=myProgramV1.0
##reference=hxb2.fas
##INFO=<ID=RC,Number=1,Type=String,Description="Reference Codon">
##INFO=<ID=AC,Number=.,Type=String,Description="Alternate Codon">
##INFO=<ID=ACF,Number=.,Type=Float,Description="Alternate Codon Frequency, for each Alternate Codon, in
    the same order as listed.">
##FILTER=<ID=af0.01,Description="Set if True; alt_freq<0.01">
#CHROM  GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE INFO
hxb2    RT      48      S       *       af0.01  0.0031    324      RC=tca;AC=tAa;ACF=0.0031
hxb2    RT      103     K       N       PASS    0.0779    154      RC=aaa;AC=aaC;ACF=0.0779
hxb2    RT      117     S       Q       af0.01  0.0033    299      RC=tca;AC=CAa;ACF=0.0033
hxb2    RT      118     V       F       af0.01  0.0065    306      RC=gtt;AC=Ttt;ACF=0.0065
hxb2    RT      174     Q       K       af0.01  0.0091    659      RC=caa;AC=Aaa;ACF=0.0091
hxb2    RT      212     W       G       af0.01  0.0044    1133     RC=tgg;AC=Ggg;ACF=0.0044
hxb2    RT      248     E       K       af0.01  0.0022    1394     RC=gaa;AC=Aaa;ACF=0.0022
```

1.2 Meta-information lines

File meta-information is included after the ## string and must be key=value pairs. It is strongly encouraged that information lines describing the INFO and FILTER entries used in the body of the AAVF file be included in the meta-information section. Although they are optional, if these lines are present then they must be completely well-formed.

1.2.1 File format

A single 'fileformat' field is always required, must be the first line in the file, and details the AAVF format version number. For example, for AAVF version 1.0, this line should read:

```
##fileformat=AAVFv1.0
```

1.2.2 Information field format

INFO fields should be described as follows (first four keys are required, source and version are recommended):

```
##INFO=<ID=ID,Number=number,Type=type,Description="description",Source="source",Version="version">
```

Possible Types for INFO fields are: Integer, Float, Flag, Character, and String. The Number entry is an integer that describes the number of values that can be included with the INFO field. For example, if the INFO field contains a single number, then this value should be 1; if the INFO field describes a pair of numbers, then this value should be 2 and so on. If the number of possible values varies, is unknown, or is unbounded, then this value should be '.'.

The 'Flag' type indicates that the INFO field does not contain a Value entry, and hence the Number should be 0 in this case. The Description value must be surrounded by double-quotes. The double-quote character can be escaped with '"' and the backslash character with '\'. Source and Version values likewise should be surrounded by double-quotes and specify the annotation source (case-insenstive, e.g. "sdrm") and exact version (e.g. "2009"), respectively for computational use.

1.2.3 Filter field format

FILTERs that have been applied to the data should be described as follows:

```
##FILTER=<ID=ID,Description="description">
```

1.3 Header line syntax

The header line names the 9 fixed, mandatory columns. These columns are as follows:

1. #CHROM
2. GENE
3. POS
4. REF
5. ALT
6. FILTER
7. ALT_FREQ
8. COVERAGE
9. INFO

1.4 Data lines

1.4.1 Fixed lines

There are 9 fixed fields per record. All data lines are tab-delimited. In all cases, missing values are specified with a dot ('.'). Fixed fields are:

1. CHROM - chromosome: An identifier from the reference genome. All entries for a specific CHROM should form a contiguous block within the AAVF file. The colon symbol (:) must be absent from all chromosome names to avoid parsing errors when dealing with breakends. (String, no white-space permitted, Required)
2. GENE - gene: An identifier for a coding sequence within the CHROM. All entries for a specific GENE should form a contiguous block within the AAVF file. The colon symbol (:) must be absent from all chromosome names to avoid parsing errors when dealing with breakends. (String, no white-space permitted, Required)
3. POS - position: The reference position within the gene specified, with 1st amino acid in the gene having position 1. Positions are sorted numerically, in increasing order, within each GENE sequence. It is permitted to have multiple records with the same POS. (Integer, Required)
4. REF - reference amino acid(s): Each amino acid must be one of A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,X,* (case insensitive). Multiple amino acids are permitted. The value in the POS fields refers to the position of the first amino acid in the String. For simple insertions and deletions in which either the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT Strings must include the amino acid before the event (which must be reflected in the POS field), unless the event occurs at position 1 on the contig in which case it must include the amino after the event. (String, Required)
5. ALT - alternate amino acid(s): Each amino acid must be one of A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y,X,* (case insensitive), where 'X' represents an ambiguous amino acid and '*' represents a stop amino acid. (String, Required)
6. FILTER - filter status: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise if the site has not passed all filters, a semicolon-separated list of codes for filters that fail. e.g. "af0.01" indicates that at this site the ALT_FREQ is below 0.01. '0' is reserved and should not be used as filter String. If filters have been applied, then this field should be set to the missing value. (String, no white-space or semi-colons permiited)
7. ALT_FREQ - alternate amino acid frequency: Frequency of the alternate allele. (Float, Required)
8. COVERAGE - coverage at that position: Number of reads that cover the POS. (Integer, Required)
9. INFO - additional information: (String, no white-space, semi-colons, or equals-signs permitted; commas are permitted only as delimiters for list of values) INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: `<key>=<data>[,data]`. Arbitrary keys are permitted, although the following sub-fields are reserved (albeit optional):

- RC: reference codon, the codon that makes up the REF amino acid(s).
- AC: alternate codon, the codon that makes up the ALT amino acid(s).
- ACC: alternate codon count (number of reads containing that codon) for each alternate codon, in the same order as listed
- ACF: alternate codon frequency, for each alternate codon, in the same order as listed

2 Understanding the AAVF format

AAVF records use a single general system for representing genetic variation data composed of:

- Allele: representing single genetic haplotypes
- AAVF record: a record holding all the segregating alleles at a locus

AAVF records use a simple haplotype representation for REF and ALT alleles to describe variant haplotypes at a locus. ALT haplotypes are constructed from the REF haplotype by taking the REF allele amino acids at the POS in the gene within the reference genotype and replacing them with the ALT amino acids. In essence, the AAVF record specifies a-REF-t and the alternative haplotypes are a-ALT-t for each alternative allele.

3 Representing variation in AAVF records

3.1 Creating AAVF entries for Synonymous and Non-synonymous mutations

3.1.1 Example 1

For example, suppose we are looking at a locus within the **a** gene in the **my__chrom** genome:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l K k s	gga ctc AAG aaa tcc	K has a silent mutation w.r.t. to the reference sequence
2	g l N k s	gga ctc AAT aaa tcc	K amino acid is a N, in a portion of the viri

Representing these as AAVF records would be done as follows:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE  INFO
my_chrom  a       3       K       K       PASS    0.95      1000      RC=aaa;AC=aaa,aaG;ACF=0.75,0.20
my_chrom  a       3       K       N       PASS    0.05      1000      RC=aaa;AC=aaT;ACF=0.05
```

3.2 Decoding AAVF entries for Synonymous and Non-synonymous mutations

3.2.1 Synonymous mutation AAVF record

Suppose I received the following AAVF record:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE  INFO
my_chrom  a       2       L       L       PASS    1.0       1000      RC=ctc;AC=ctc,ctT;ACF=0.75,0.25
```

This is a synonymous mutation since the alt amino acid is the same as the reference amino acid, and the 'AC' INFO field contains a codon which is difference from the reference codon, so I have the two following haplotypes:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g L k k s	gga CTC aaa aaa tcc	L is the reference amino acid
1	g L K k s	gga CTT aaa aaa tcc	L has a silent mutation w.r.t. to the reference sequence

3.2.3 Non-synonymous mutation AAVF record

Suppose I received the following AAVF record:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       4       K       I       PASS    0.75      1000     RC=aaa;AC=aTa;ACF=0.75
```

This is a non-synonymous mutation since the alt amino acid differs from the reference amino acid, so I have the two following haplotypes:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l k K s	gga ctc aaa AAA tcc	K is the reference amino acid
1	g l k I s	gga ctc aaa ATA tcc	K amino acid is a I, in a portion of the virus population

3.3 Creating AAVF entries for Insertions and Deletions

3.3.1 Example 1

For example, suppose we are looking at a locus with the **a** gene in the **my_chrom** genome:

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l - k s	gga ctc — aaa tcc	K amino acid is deleted w.r.t. to the reference sequence
2	g l KKk s	gga ctc AAAAAA aaa tcc	K amino acid is inserted w.r.t. to the reference sequence

Representing these as AAVF records would be done as follows:

1. A single amino acid deletion of K at position 3 becomes REF=LK, ALT=L
2. A single amino acid insertion of K after position 3 becomes REF=K, ALT=KK

Note: that the positions must be sorted in increasing order:

```
#CHROM    GENE    POS     REF     ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       2       LK      L       PASS    0.5       1000     RC=ctcaaa;AC=ctc
my_chrom  a       3       K       KK      PASS    0.5       1000     RC=aaa;AC=aaaaaa;ACF=0.5
```

3.4 Decoding AAVF entries for Insertions and Deletions

3.4.1 Insertion AAVF record

Supposed I receive the following AAVF record:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       3       K      KK      PASS    0.5       1000     RC=aaa;AC=aaaaaa;ACF=0.5
```

This is an insertion since the reference amino acid K is being replaced by K [the reference amino acid] plus one insertion amino acid K in such a way that a gap is opened in the reference. Again there are only two alleles so I have the two following segrating haplotypes.

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA — aaa tcc	K is the reference amino acid
1	g l KKk s	gga ctc AAA AAA aaa tcc	K amino acid is inserted w.r.t. to the reference sequence

3.4.2 Deletion AAVF record

Supposed I receive the following AAVF record:

```
#CHROM    GENE    POS     REF    ALT     FILTER  ALT_FREQ  COVERAGE INFO
my_chrom  a       2       LK     L       PASS    0.5       1000     RC=ctcaaa;AC=ctc
```

This is a deletion of one reference amino acid since the reference allele LK is being replaced by just the L [the reference amino acid]. Again there are only two alleles so I have the two following seggregating haplotypes.

Example	Amino Acid Sequence	Nucleotide Sequence	Alteration
Ref	g l K k s	gga ctc AAA aaa tcc	K is the reference amino acid
1	g l - k s	gga ctc — aaa tcc	K amino acid is deleted w.r.t. to the reference sequence