# quasitools: A Collection of Tools for Viral Quasispecies Analysis

Eric Marinier[1], Eric Enns[1], Camy Tran[1], Matthew Fogel[1], Cole Peters[1], Ahmed Kidwai[1], Hezhao Ji[2,3], Gary Van Domselaar[1,3]

**1 Bioinformatics Core at the National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Manitoba, Canada**
**2 National HIV and Retrovirology Laboratories at JC Wilt Infectious Diseases Research Center, Public Health Agency of Canada, Winnipeg, Manitoba, Canada**
**3 Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Manitoba, Canada**

## Abstract

**Summary:** quasitools is a collection of newly-developed, open-source tools for analyzing viral quasispcies data. The application suite includes tools with the ability to create consensus sequences, call nucleotide, codon, and amino acid variants, calculate the complexity of a quasispecies, and measure the genetic distance between two similar quasispecies. These tools may be run independently or in user-created workflows.

**Availability:** The quasitools suite is a freely available application licensed under the Apache License, Version 2.0. The source code, documentation, and file specifications are available at: https://phac-nml.github.io/quasitools/

**Contact:** gary.vandomselaar@canada.ca

## Introduction

The existing predominance and regular emergence of viral pathogens represent a significant public health threat worldwide. The ability to rapidly identify and characterize viral disease outbreaks is an essential component of public health response. Many effective conventional molecular methods exist and are routinely applied for viral disease detection, surveillance, and characterization. However, next-generation sequencing (NGS) technologies offer a far more sensitive, accurate, and comprehensive means for studying, characterizing, and monitoring viral pathogens.

Viral quasispecies are viruses that replicate in high numbers but with low fidelity, which results in a complex and dynamic spectrum of mutations within an infected individual [1]. Analysis of viral quasispecies poses a unique computational challenge, owing to the unprecedented data volume and complexity of viral quasispecies, sequence variations at diverse frequencies across the genome, demanding quality control strategies required for NGS data processing, and a lack of consensus in performing such analyses. Bioinformatics efforts to address the unique challenges of virology thus far have been underwhelming relative to their biomedical value and public health importance [2].

Nevertheless, there are a number of tools that exist for analyzing viral quasispecies. There are many tools that estimate the global diversity of viral quasispecies using read graph-based methods [3–7], probabalistic clustering methods [8–10], and *de novo* assembly methods [11]. HIVE [12] is a web environment that contains a tool for population analysis of viral quasispecies. The tool correlates distant mutations, identifies clones of low coverage, and outputs diagrams visualizing this information. PAQ [13] is capable of partitioning quasispecies sequences into groups that are genetically similar. Beyond these examples, there are numerous pipelines for variant calling and drug resistance identification [14–20]. Many applications designed for quasispecies analysis deal with specific issues in resolving viral quasispecies. The motivation behind quasitools is to provide a single collection of open-source, general-purpose

| Tool | Description |
|---|---|
| aacoverage | builds an amino acid consensus and returns its coverage |
| call aavr | calls amino acid mutations for a BAM file and a supplied reference file |
| call codonvar | calls codon variants for a BAM file and a supplied reference file |
| call ntvar | calls nucleotide variants for a BAM file and a supplied reference file |
| complexity | reports the complexity of a quasispecies using several measures |
| consensus | generates a consensus sequence from a BAM file |
| distance | measures the distance between two quasispecies using angular cosine distance |
| dnds | calculates the dN/dS value for each region in a BED file |
| drmutations | identifies AA mutations corresponding to known drug resistance mutations |
| hydra | identifies HIV drug resistance mutations in an NGS dataset |
| quality | performs quality control on FASTQ reads |

Table 1: An overview of the tools available in quasitools.

tools for quasispecies analysis that are designed to operate seamlessly with each other.

## Implementation

quasitools is a Python application that provides several tools for quasispecies analysis, including variant calling, consensus generation, and reporting of viral quasispecies complexity (Table 1). These tools may be run independently or together in user-created workflows. Here we provide a brief summary of some of these tools.

The *quality* tool performs basic quality control on read sequencing data. It can filter FASTQ-formatted reads based on average quality score, median quality score, and read length. It can also mask low-quality bases and perform iterative read trimming. The *consensus* tool generates the consensus sequence of a quasispecies, using either a most-frequent base strategy or using a strategy that reports ambiguous IUPAC bases at positions where there is insufficient agreement.

There are multiple variant calling tools that can identify nucleotide (*call ntvar*), codon (*call codonvar*), and amino acid variants (*call aavar*) within a quasispecies, given a BAM alignment file and a FASTA reference file. The output of these variant calling tools may be provided to other tools in order to identify drug-resistant mutations.

We have developed a tool, *complexity*, which implements a wide variety of quasispecies complexity measurements for haplotypes, as described by Josep et al. 2016 [1]. These multidimensional measures capture the number of mutants, frequency of different haplotypes, and viral population size. Additionally, we have extended this framework for evaluating quasispecies complexity from amplicons to $k$-mers across a sequence pileup. This enhancement enables the study of quasispecies complexity across the genome using NGS data, but is limited by the length and accuracy of the sequencing reads.

We provide a tool, *distance*, for estimating the genetic distance between two quasispecies, which may be a useful and realtively quickly derived approximation for evolutionary relatedness among quasispecies of the same species. The tool represents each quasispecies pileup as a vector and calculates the angular cosine distance between each of these vectors. We use angular cosine distance as it is accommodates vectors that are both sparse and varied in their depth of coverage. Furthermore, the user may normalize the pileup vectors to prevent bias in relative coverage within a single pileup, where a region with exceptional coverage would dominate the distance calculation.

HyDRA is an annotated, reference-based pipeline, which analyzes NGS data for genotyping HIV-1 drug resistance mutations. HyDRA combines the analyses of many tools in quasitools in order to identify drug resistance mutations in HIV quasispecies. It uses an annotated HXB2

sequence for reference mapping and stringent variant calling to identify HIV drug resistant mutations based on the Stanford HIV Drug Resistance Database and the 2009 WHO list for Surveillance of Transmitted HIV Drug Resistance. As HyDRA is built using the tools available within quasitools, it is possible for users to use quasitools to create similar workflows for other viral quasispecies.

The quasitools package is freely available as source code, as a Conda package, and individually as tools on the public Galaxy Tool Shed for use on the Galaxy web-based platform.

*Funding:* This work was funded by a grant from the Canadian Federal Government Genomics Research and Development Initiative (GRDI).

*Conflict of Interest:* None declared.

## References

[1] Josep Gregori, Celia Perales, Francisco Rodriguez-Frias, Juan I Esteban, Josep Quer, and Esteban Domingo. Viral quasispecies complexity measures. *Virology*, 493:227–237, 2016.

[2] Martin Hölzer and Manja Marz. Software dedicated to virus sequence analysis "bioinformatics goes viral". In *Advances in Virus Research*, volume 99, pages 233–257. Elsevier, 2017.

[3] Mattia CF Prosperi and Marco Salemi. QuRe: Software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1):132–133, 2011.

[4] Osvaldo Zagordi, Arnab Bhattacharya, Nicholas Eriksson, and Niko Beerenwinkel. ShoRAH: Estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12(1):119, 2011.

[5] Irina Astrovskaya, Bassam Tork, Serghei Mangul, Kelly Westbrooks, Ion Măndoiu, Peter Balfe, and Alex Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. In *BMC Bioinformatics*, volume 12, page S1. BioMed Central, 2011.

[6] Nicholas Mancuso, Bassam Tork, Pavel Skums, Ion Măndoiu, and Alex Zelikovsky. Viral quasispecies reconstruction from amplicon 454 pyrosequencing reads. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 94–101. IEEE, 2011.

[7] Shawn T O'Neil and Scott J Emrich. Haplotype and minimum-chimerism consensus determination using short sequence data. *BMC Genomics*, 13(2):S4, 2012.

[8] Christopher Quince, Anders Lanzen, Russell J Davenport, and Peter J Turnbaugh. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12(1):38, 2011.

[9] Sandhya Prabhakaran, Melanie Rey, Osvaldo Zagordi, Niko Beerenwinkel, and Volker Roth. HIV-haplotype inference using a constraint-based dirichlet process mixture model. In *Machine Learning in Computational Biology (MLCB) NIPS Workshop*, pages 1–4, 2010.

[10] Armin Töpfer, Osvaldo Zagordi, Sandhya Prabhakaran, Volker Roth, Eran Halperin, and Niko Beerenwinkel. Probabilistic inference of viral quasispecies subject to recombination. *Journal of Computational Biology*, 20(2):113–123, 2013.

[11] Franklin Bristow. De novo sequence assembly of viral quasispecies. *University of Manitoba*, 2012.

[12] Vahan Simonyan and Raja Mazumder. High-performance integrated virtual environment (HIVE) tools and applications for big data analysis. *Genes*, 5(4):957–981, 2014.

[13] Prasith Baccam, Robert J. Thompson, Olivier Fedrigo, Susan Carpenter, and James L. Cornette. PAQ: Partition Analysis of Quasispecies . *Bioinformatics*, 17(1):16–22, 01 2001.

[14] Mark Howison, Mia Coetzer, and Rami Kantor. Measurement error and variant-calling in deep illumina sequencing of HIV. *bioRxiv*, page 276576, 2018.

[15] Xiao Yang, Patrick Charlebois, Alex Macalalad, Matthew R Henn, and Michael C Zody. V-Phaser 2: Variant inference for viral populations. *BMC Genomics*, 14(1):674, 2013.

[16] Bie MP Verbist, Kim Thys, Joke Reumers, Yves Wetzels, Koen Van der Borght, Willem Talloen, Jeroen Aerssens, Lieven Clement, and Olivier Thas. VirVarSeq: A low-frequency virus variant detection pipeline for illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics*, 31(1):94–101, 2014.

[17] Michael Huber, Karin J Metzner, Fabienne D Geissberger, Cyril Shah, Christine Leemann, Thomas Klimkait, Jürg Böni, Alexandra Trkola, and Osvaldo Zagordi. MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *Journal of Virological Methods*, 240:7–13, 2017.

[18] Matthias Döring, Joachim Büch, Georg Friedrich, Alejandro Pironti, Prabhav Kalaghatgi, Elena Knops, Eva Heger, Martin Obermeier, Martin Däumer, Alexander Thielen, et al. geno2pheno [ngs-freq]: A genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Research*, 46(W1):W271–W277, 2018.

[19] Ana Garcia-Diaz, Adele McCormick, Clare Booth, Dimitri Gonzalez, Chalom Sayada, Tanzina Haque,

Margaret Johnson, and Daniel Webster. Analysis of transmitted HIV-1 drug resistance using 454 ultra-deep-sequencing and the deepchek®-hiv system. *Journal of the International AIDS Society*, 17:19752, 2014.

[20] Hezhao Ji, Eric Enns, Chanson J Brumme, Neil Parkin, Mark Howison, Emma R Lee, Rupert Capina, Eric Marinier, Santiago Avila-Rios, Paul Sandstrom, et al. Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: the Winnipeg Consensus. *Journal of the International AIDS Society*, 21(10):e25193, 2018.