8-10-2021

Methods for Viral Intra-Host and Inter-Host Data Analysis for Next-Generation Sequencing Technologies

Sergey Knyazev

METHODS FOR VIRAL INTRA-HOST AND INTER-HOST DATA ANALYSIS FOR NEXT-GENERATION SEQUENCING TECHNOLOGIES

by

Sergey Knyazev

Under the Direction of Alex Zelikovsky, PhD

ABSTRACT

The deep coverage offered by next-generation sequencing (NGS) technology has facilitated the reconstruction of intra-host RNA viral populations at an unprecedented level of detail. However, NGS data requires sophisticated analysis dealing with millions of error-prone short reads. This dissertation will first review the challenges and methods for viral NGS genomic data analysis in the NGS era. Second, it presents a software tool CliqueSNV for inferring viral quasispecies based on extracting pairs of statistically linked mutations from noisy reads, which effectively reduces sequencing noise and enables identifying minority haplotypes with a frequency below the sequencing error rate. Finally, the dissertation describes algorithms VOICE and MinDistB for inference of relatedness between viral samples, identification of transmission clusters, and sources of infection.

INDEX WORDS: Algorithms, intra-host and inter-host viral populations, viral genome assembly, viral haplotype and mutation calling, outbreak investigation, next-generation sequencing

METHODS FOR VIRAL INTRA-HOST AND INTER-HOST DATA ANALYSIS FOR NEXT-GENERATION SEQUENCING TECHNOLOGIES

by

Sergey Knyazev

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2021

METHODS FOR VIRAL INTRA-HOST AND INTER-HOST DATA ANALYSIS FOR NEXT-GENERATION SEQUENCING TECHNOLOGIES

by

Sergey Knyazev

Committee Chair: Alex Zelikovsky

Committee: Pavel Skums

Robert Harrison

William M. Switzer

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2021

DEDICATION

To Ekaterina and Lev

ACKNOWLEDGMENTS

I am grateful to my scientific adviser Dr. Alex Zelikovsky for his care and guidance during my Ph.D. journey and helping me to grow as a professional, a collaborator, and a person. I am thankful to my mentors at CDC Bill Switzer and Ells Campbell who helped me to apply my bioinformatic tools toward public health tasks. I appreciate the help of my business mentors Dr. Stas Samarin and Dr. Andrew Slivker who helped with discovering of a business potential of my scientific projects. I acknowledge Dr. Serghei Mangul, Dr. Pavel Skums, Dr. Bogdan Pasaniuc, Dr. Robert Harrison, Dr. Murray Patterson, Dr. Frank Stewart, and Dr. Ion Mandoiu for their advises and consultations. It was a pleasure to work with my peers Dr. Olga Glebova, Dr. Ekaterina Gerasimov, Dr. Alexander Artyomenko, Dr. Igor Mandric, Dr. Andrew Melnyk, Dr. Pelin Icer, Dr. Mark Grinshpon, Viachaslau Tsyvina, Daniel Novikov, Lauren Hughes, Fil Rondel, Bikram Sahoo, Roya Hosseini, Fatemeh Mohebbi and master and undegraduate students from GSU, UCLA, and USC labs who worked and collaborated with me.

Chapter 1 and 2, is based on the material as it appears in S. Knyazev, L. Hughes, P. Skums, A. Zelikovsky, ”Epidemiological data analysis of viral quasispecies in the next-generation sequencing era,” Briefings in Bioinformatics 22(1):96-108, 2021. The dissertation author was the lead author of this paper.

Chapter 3, is based on the material as it appears in S. Knyazev, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, Y. Porozov, E. Campbell, W. Switzer, P. Skums, A. Zelikovsky, Efficient Noise Reduction Technique forSensitive Assembly and Drug-ResistanceDetection from Viral NGS

Data. Nuclear Acids Research 2021 Jul 2:gkab576. doi: 10.1093/nar/gkab576. Epub ahead of print. PMID: 34214168. The dissertation author was the primary developer of the CliqueSNV project and one of the two lead authors of this paper.

Chapter 4, is based on the material as it appears in O.Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, P. Skums, ”Inference of genetic relatedness between viral quasispecies from sequencing data,” BMC Genomics, 18(Suppl. 10):918,2017. The dissertation author was the developer of the VOICE tool in the project and one of the two lead authors of this paper.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . **v**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **xi**

1 INTRODUCTION **1**

1.1 RNA viruses and viral intra-host and inter-host populations 1

1.2 Application of next-generation sequencing for viral studies 2

1.3 Problem formulations 4

1.4 Contributions 6

1.5 Refereed Journal Articles 7

1.6 Refereed Articles in Conference Proceedings 10

2 EPIDEMIOLOGICAL DATA ANALYSIS OF VIRAL QUASISPECIES IN THE NEXT-GENERATION SEQUENCING ERA **14**

2.1 Viral populations analysis problem and challenges 14

2.2 The primary analysis of viral next-generation sequencing data 15

2.3 Basic primary analysis 15

2.3.1 Single nucleotide variant calling 17

2.3.2 Viral haplotype variant calling 20

2.4 Secondary analysis of viral next-generation sequencing data 24

2.4.1 Predicting drug resistance 24

2.4.2 Estimating infection recency 28

2.4.3 Outbreak investigation 29

2.5 Molecular surveillance systems and databases 34

3 CLIQUESNV - A METHOD FOR INFERRING VIRAL QUASISPECIES USING NGS 36

3.1 CliqueSNV algorithm 39

3.2 Results 46

3.2.1 Intra-host viral population sequencing benchmarks 46

3.2.2 Validation metrics for viral population inference 49

3.2.3 Performance of haplotyping methods 51

3.2.4 Runtime comparison 58

3.3 Discussion 61

4 INTER-HOST VIRAL ANALYSIS USING NGS 64

4.1 Introduction 64

4.2 Methods 66

4.2.1 Viral outbreak inference ($VOICE$) simulation method 66

4.2.2 MinDistB method 71

4.3 Results 72

4.3.1 Data sets 73

4.3.2 Prediction of epidemiological characteristics 73

4.4 Conclusions 77

REFERENCES 78

LIST OF TABLES

Table 2.1 SNV calling software tools for viral NGS data 19

Table 2.2 Haplotype calling software tools for viral NGS data 22

Table 2.3 Detection of drug-resistant mutations in clinical studies: NGS vs Sanger sequencing 27

Table 2.4 Outbreak investigation software tools for viral NGS data 33

Table 3.1 Four experimental and two simulated sequencing datasets of human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). The datasets contain MiSeq and PacBio reads from intra-host viral populations consisting of two to ten variants each with frequencies in the range of 0.1-50%, and Hamming distances between variants in the range of 0.1-3.5%. 46

Table 3.2 Prediction statistics of haplotype reconstruction methods using experimental and simulated (a) MiSeq and (b) PacBio data. The precision and recall was evaluated stringently such that if a predicted haplotype has at least one mismatch to its closest answer, then that haplotype is scored as a false positive. 52

Table 3.3 Earth Movers' Distance from predicted haplotypes to the true haplotype population and haplotyping method improvement. Four haplotyping methods(aBayesQR, CliequeSNV, Consensus, PredictHaplo) are benchmarked on five MiSeq datasets (a) and IAV10exp dataset (b). The improvement shows how much better is prediction of haplotyping method over inferred consensus, and it is calculated as $\frac{(EMD_c - EMD_m) \times 100\%}{EMD_c}$, where EMD_c is an EMD for consensus, and EMD_m is an EMD for method. CliqueSNV outperformed all other methods in accuracy on all datasets. 56

Table 3.4 Comparison of CliqueSNV with PredictHaplo and aBayesQR on simulated and real Illumina data 57

Table 3.5 Comparison of CliqueSNV with PredictHaplo and 2SNV on IAV10exp . . . 58

Table 3.6 Comparison of CliqueSNV, 2SNV and PredictHaplo on full and sub-sampled data ***(PacBio, experimental)***. For all 33.5K reads, the sign "✓" (respectively, "×") denotes fully matched (respectively, unmatched) true variant and the column FP reports the number of incorrectly predicted variants (false positives) and their total frequency. For each sub-sample size (16K,...,4K), the table reports the percent of runs when a variant is completely matched and its average frequency over runs when the variant was detected. Similarly, the column FP reports the average number of false positive variants and their average total frequency. Colors indicate the percent of matched variants: green - high percent, red - low percent. 59

Table 3.7 Running time of performed experiments (seconds) for full-length benchmarks. 60

Table 4.1 Validation results 76

LIST OF FIGURES

Figure 1.1 A molecular surveillance pipeline for software tools for primary and secondary viral NGS data analysis. 5

Figure 3.1 Schematic representation of the CliqueSNV algorithm, where SNV is single nucleotide variation. 41

Figure 3.2 A typical distribution of errors in PacBio reads. The heavy tail indicates that a significant portion of errors is accumulated by a relatively small number of reads. 44

Figure 3.3 The clique graph C_G with 5 vertice corresponding to cliques in G, 4 edges and two forbidden pairs (q_1, q_2) and (q_2, q_3). There 3 maximal connected subgraphs avoiding forbidden pairs: $\{q_1, q_4\}$ $\{q_4, q_2, q_5\}$ $\{q_5, q_3\}$ 45

Figure 3.4 Pairwise hamming distances between variants in datasets HIV9exp, HIV2exp, HIV5exp, HIV7sim, IAV10sim, and IAV10exp. 47

Figure 3.5 The number of true and false predicted haplotypes depending on the number of accepted mismatches for five benchmarks: (A) HIV9exp; (B) HIV2exp; (C) HIV5exp; (D) HIV7sim; (E) IAV10sim. Two haplotypes are regarded identical if the Hamming distance between them is at most the number of accepted mismatches. 53

Figure 3.6 Matching distances $E_{T\leftarrow P}$ and $E_{T\rightarrow P}$ between a true haplotype population and a reconstructed haplotype population for five benchmark datasets for human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). Matching distance $E_{T\leftarrow P}$ is shown on the x-axis and $E_{T\rightarrow P}$ is shown on the y-axis for each benchmark. Smaller matching distances indicate better approximation of a true haplotype population T by a reconstructed haplotype population P. Haplotype populations were reconstructed with CliqueSNV, aBayesQR, PredictHaplo and a population consisting of a single consensus haplotype. 54

Figure 3.7 Earth Movers' Distance (EMD) between true and reconstructed haplotype populations. Four haplotyping methods (CliqueSNV, aBayesQR, PredictHaplo, Consensus) are benchmarked using three experimental and two simulated datasets for human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). For all benchmarks the CliqueSNV predictions are the closest to the true populations. 55

Figure 3.8 The number of reads assigned to different number of cliques in HIV Illumina dataset. 58

Figure 3.9 Runtime of PredictHaplo (PH), 2SNV and CliqueSNV on datasets with different sizes. 60

Figure 3.10 CliqueSNV runtime on datasets with different reference length and same coverage (about 1M reads in total). 61

Figure 4.1 Edge subdividing 67

Figure 4.2 All possible moves of a vertex v 69

Figure 4.3 δ-Crossing between two viral populations P_1 and P_2 $l \leq d(u,v)+\delta$; (A) $|B_\delta| = 5$; (B) $|B_\delta| = 2$ 70

Figure 4.4 Intuition behind the MinDistB method. (A) Related samples – crossing is between old survived variants (B) Unrelated samples –crossing is between many young variants which are close to each other by chance. 70

Figure 4.5 ROC curve for pairs relatedness detection 76

CHAPTER 1
INTRODUCTION

1.1 RNA viruses and viral intra-host and inter-host populations

A virus is a submicroscopic entity that intrudes a living cell and exploits the cell's resources to replicate itself. The virus, like the cell, uses genetic information to pass data from a generation to the next generation. The major difference between the virus and the cell is that the virus cannot replicate itself without a host because the viral genome doesn't carry all the information required for the replication. The virus carries only supplementary information that is enough to add to the cell to force the cell to produce viral clones. Some viruses are pathogens because they disrupt the cell's life balance.

The class of viruses that use RNA to carry genetic information are called RNA viruses[172]. Viral RNA can be either single-stranded or double-stranded. RNA viruses cause deceases such as the common cold, influenza, COVID-19, SARS, HIV, hepatitis, Ebola, rabies, polio, and measles.

Due to error-prone replication, RNA viruses mutate at rates estimated to be as high as 10^{-3} substitutions per nucleotide per replication cycle[51]. Since mutations are generally well tolerated, such viruses exist in infected hosts as "quasispecies" - a term used by virologists to describe populations of closely related genomic variants[45,46,52,115]. Genetic heterogeneity of viral quasispecies has major biological implications, contributing to the efficiency of virus transmission, tissue tropism, virulence, disease progression, and the emergence of drug/vaccine-resistant variants[18,50,65,79,144].

1.2 Application of next-generation sequencing for viral studies

With the advent of next-generation sequencing (NGS) technologies, molecular epidemiology and virology are undergoing a fundamental transformation that promises to revolutionize our approach to epidemiological data analysis, disease prevention, and treatment[33,39,71,131]. NGS has already shown its potential to advance epidemiological practices and it is steadily moving into clinical practices. There are numerous examples of successful applications of NGS for studying viruses such as coronavirus[154], influenza[161,116,173,150,168,58], HIV[86,38,23,36,76,59], Hepatitis[163,176,27,64], Ebola[136,77], and Zika[137].

NGS allows sequencing with the unprecedently deep coverage, which is crucial for characterizing intra-host viral population complexity. However, inferring and analyzing the viral population from NGS data is computationally challenging and requires specialized, highly sophisticated computational tools[132]. Even for NGS technologies offering very deep coverage, the presence of sequencing errors makes it difficult to distinguish between rare variants and sequencing errors. Additionally, low intra-host viral diversity complicates assembling whole-genome sequences that are necessary for the unique identification of viral haplotypes. Therefore, the analysis of heterogeneous virus populations complemented by technological developments.

The viral population reconstructed from NGS data can be further used for the detection of drug resistance in the patients' samples as well as the age of infection. The importance of this detection is constantly growing[117], especially for Influenza[130], HCV[106], and HIV[21,177] because of the high prevalence of these diseases in the population. As for HIV, there is an additional problem. Since HIV has no cure, its treatment can only slow down its progression, and the development of drug

resistance creates the risk of losing a drug forever as a treatment option for the patient. This is further complicated by the increasing longevity of HIV patients and the prevalence of the disease among the general population. Since viruses exist as a swarm of haplotypes, it is crucial to detect minority drug-resistant populations.

The haplotypes inferred from NGS data can also be very effective for outbreak investigation. Millions of viral variants that are carried in the samples of thousands of infected individuals can be analyzed with the help of NGS. Molecular data collected from densely sampled outbreaks in large high-risk communities are of particular interest since it allows for the first time to study the evolution of heterogeneous intra-host viral populations within a single evolutionary space under frequent transmissions between hosts[160,70,118]. The growing knowledge about social network structures and progress in the development of methods for the collection of large volumes of socio-behavioral and geographic data gives us new information about the conditions of disease spread[26,129,101]. The availability of such large-scale datasets provides a new opportunity to implement massive molecular surveillance and forecasting of viral diseases[142,97,105,1,99,24]. Deployment of massive molecular surveillance programs intends to facilitate our understanding of virus evolution, enabling the development of more effective public health intervention strategies. To be effective, molecular surveillance and forecasting should analyze unprecedented amounts of heterogeneous biomedical data. This requires extensive computational methods for processing, integrating and analyzing big data that is both epidemiological and molecular. In addition, this requires new mathematical models that allow for describing, understanding and predicting complex multidimensional-linear disease dynamics.

The remainder of the review will discuss the pipeline of software tools for primary and secondary NGS data analysis constituting a sequencing-based molecular surveillance system (see Figure 1.1). The primary NGS data analysis consists of error correction, consensus assembly/selection, read alignment, and inference of intra-host viral population including SNV calling and haplotype reconstruction. The secondary NGS data analysis includes intra-host analysis such as detection of drug resistance and estimations of the age of infection as well as inter-host analysis such as outbreak detection and investigation. Finally, we review existing molecular surveillance systems that integrate all the above analyses.

1.3 Problem formulations

This dissertation addresses the following problems:

- Given NGS reads from DNA/RNA intra-host viral sample, reconstruct intra-host viral population, i.e. all distinct viral variants (haplotypes) and their frequencies.
- Given NGS reads from DNA/RNA intra-host viral sample, reconstruct intra-host viral single nucleotide variants (SNVs), i.e. all distinct SNVs and their frequencies.
- Given haplotypes from two intra-host viral populations A and B, decide whether
  (i) A and B are related
  (ii) A infected B or B infected A
- Given haplotypes from a set of intra-host viral populations, find
  (i) the source of an outbreak

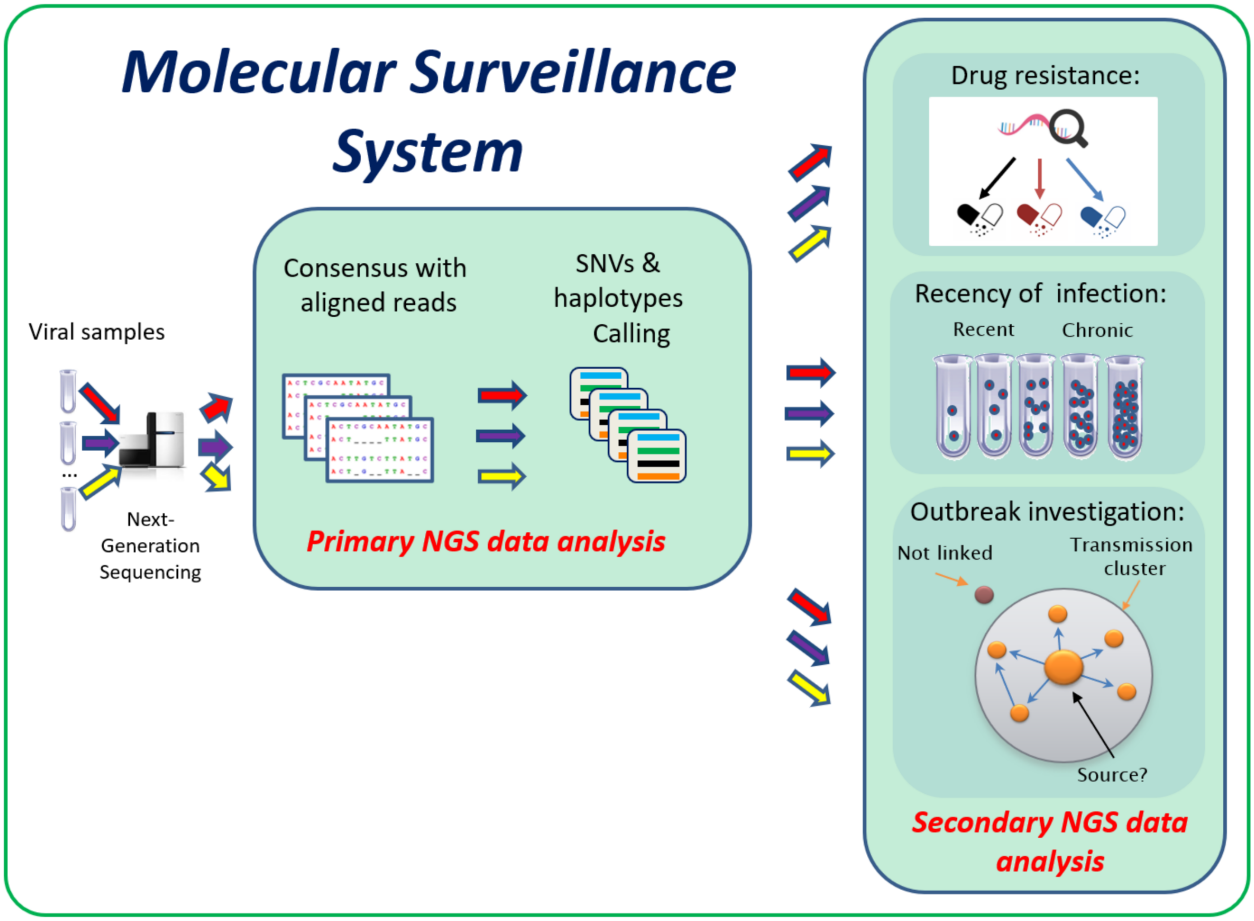

Figure 1.1 A molecular surveillance pipeline for software tools for primary and secondary viral NGS data analysis.

(ii) the transmission clusters corresponding to individual outbreaks

- Given:

(i) real sequencing benchmark, including reads and ground truth haplotype population

(ii) parameters for simulation such as error rate, coverage, average distance between haplotypes

Design a set of new benchmarks with given parameters mimicking given real sequencing

benchmark.

1.4 Contributions

The dissertation describes the following contributions:

- Summarized the state-of-the-art tools created for NGS data analysis for viral quasispecies. Gathered tools for primary data analysis for tasks of NGS error correction, SNV calling, and haplotype calling. Gathered tools for secondary data analysis for tasks of drug-resistance detection, estimating recency of infection, and outbreak investigation.

- Designing a novel haplotype assembly algorithm CliqueSNV which is based on representation of haplotype assembly as a clique enumeration problem. This approach allows efficiently cluster groups of SNVs and assign them to haplotypes. The algorithm also estimates frequencies of haplotypes by Expectation-Maximization methods, which assign sequencing reads to SNV clusters. CliqueSNV is more accurate than other methods that was proven on a series of real sequencing benchmarks.

- Two novel viral outbreak investigation tools VOICE and MinDistB that allow determine the relatedness between viral samples, source of infections, and the direction of viral spread. VOICE uses Markov process simulation to reconstruct the process of viral evolution in a space of observed viral haplotypes. MinDistB is improved version of MinDist[30] with improved sensitivity and specificity.

- Developing benchmarks for NGS software. Created a novel approach for modifying a real sequencing benchmark for modifying benchmark ground truth and error rate. That helped

test error correction tools on wide range of settings.

1.5 Refereed Journal Articles

19. **S. Knyazev**, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, Y. Porozov, E. Campbell, W. Switzer, P. Skums, A. Zelikovsky, Efficient Noise Reduction Technique forSensitive Assembly and Drug-ResistanceDetection from Viral NGS Data. **Nuclear Acids Research** 2021 Jul 2:gkab576. doi: 10.1093/nar/gkab576. Epub ahead of print. PMID: 34214168

18. M. Alser, J. Rotman, K. Taraszka, H. Shi, P. I. Baykal, H. T. Yang, V. Xue, **S. Knyazev**, B. D Singer, B. Balliu, D. Koslicki, P. Skums, A. Zelikovsky, C. Alkan, O. Mutlu, S. Mangul, "Technology dictates algorithms: Recent developments in read alignment," **Genome Biology**, 2021, accepted

17. F. Rondel, R. Hosseini, B. Sahoo, **S. Knyazev**, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, Y. Porozov, A. Zelikovsky, "Pipeline for Analyzing Activity of Metabolic Pathways in Planktonic Communities Using Metatranscriptomic Data," **Journal of Computational Biology 28(8)**: 1-14, 2021 doi: 10.1089/cmb.2021.0053

16. E. M. Campbell, A. A. Boyles, A. Shankar, J. Kim, **S. Knyazev**, W. M. Switzer, "MicrobeTrace: Retooling Molecular Epidemiology for Rapid Public Health Response," **PLOS Computational Biology**, 2021, accepted

15. I. Alexiev, E.M. Campbell, **S. Knyazev**, Y. Pan, L. Grigorova, R. Dimitrova, A. Partsuneva, A. Gancheva, A. Kostadinova, C. Seguin-Devaux, I. Elenkov, N. Yancheva, W.M. Switzer,

”Molecular Epidemiological Analysis of the Origin and Transmission Dynamics of the HIV-1 CRF01_AE Sub-Epidemic in Bulgaria,” **Viruses**. 2021; 13(1):116

14. A. Melnyk, F. Mohebbi, **S. Knyazev**, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, M. Patterson, ”Clustering Based Identification of SARS-CoV-2 Subtypes,” In: Jha S.K., Măndoiu I., Rajasekaran S., Skums P., Zelikovsky A. (eds) Computational Advances in Bio and Medical Sciences. ICCABS 2020. **Lecture Notes in Bioinformatics 12686**. 127-141

13. **S. Knyazev**, L. Hughes, P. Skums, A. Zelikovsky, ”Epidemiological data analysis of viral quasispecies in the next-generation sequencing era,” **Briefings in Bioinformatics** 22(1):96-108, 2021

12. A. Melnyk, **S. Knyazev**, F. Vannberg, L. Bunimovich, P. Skums, A. Zelikovsky, ”Using earth mover’s distance for viral outbreak investigations,” **BMC Genomics 21(Suppl 5)**:582, 2020

11. F. Rondel, R. Hosseini, B. Sahoo, **S. Knyazev**, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, A. Zelikovsky, ”Estimating Enzyme Participation in Metabolic Pathways for Microbial Communities from RNA-seq Data,” Proc. of International Symposium on Bioinformatics Research Applications (ISBRA), 2020, **Lecture Notes in Bioinformatics 12304**, 335-343

10. K.Mitchell, J.J.Brito, I.Mandric, Q.Wu, **S.Knyazev**, S.Chang, L.S.Martin, A.Karlsberg, E.Gerasimov, R.Littman, B.L.Hill, N.C.Wu, H.Yang, K.Hsieh, L.Chen, E.Littman, T.Shabani, G.Enik, D.Yao, R.Sun, J.Schroeder, E.Eskin, A.Zelikovsky, P.Skums, M.Pop, S.Mangul: ”Benchmarking of computational error-correction methods for next-generation sequencing data,” **BCB 2020**: 63:1

9. K. Mitchell, JJ. Brito, I. Mandric, Q.Wu, **S. Knyazev**, S. Chang, LS. Martin, A. Karlsberg, E. Gerasimov, R. Littman, BL. Hill, NC. Wu, H.Yang, K. Hsieh, L. Chen, E. Littman, T. Shabani, G. Enik, D. Yao, Ren Sun, J. Schroeder, E. Eskin, A. Zelikovsky, P. Skums, M. Pop, S. Mangul, "Benchmarking of computational error correction methods for next-generation sequencing data," **Genome Biology 21:71**, (2020)

8. I. Alexiev, E.M. Campbell, **S. Knyazev**, Y. Pan, L. Grigorova, R. Dimitrova, A. Partsuneva, A. Gancheva, A. Kostadinova, C. Seguin-Devaux, W.M. Switzer, "Molecular Epidemiology of the HIV-1 Subtype B Sub-Epidemic in Bulgaria," **Viruses**. 2020; 12(4):441.

7. I. Mandric, **S. Knyazev**, A. Zelikovsky, "Repeat aware evaluation of scaffolding tools," **Bioinformatics** 34(15):2530-37, 2018.

6. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, **S. Knyazev**, I.Mandric, S. Ramachandran, D. Campo, D. Jha, L. Bunimovich, E. Costenbader, C. Sexton, S. O'Connor, G. Xia, Y. Khudyakov, "QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data," **Bioinformatics** 34(1):163-70, 2018.

5. O.Glebova, **S. Knyazev**, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, P. Skums, "Inference of genetic relatedness between viral quasispecies from sequencing data," **BMC Genomics, 18(Suppl. 10)**:918,2017.

4. I. Mandric, **S. Knyazev**, C. Padilla, F. Stewart, I. I. Măndoiu, A. Zelikovsky, "Metabolic Analysis of Metatranscriptomic Data from Planktonic Communities," **Proc. of Interna-**

tional Symposium on Bioinformatics Research Applications (ISBRA), 2017, Lecture Notes in Bioinformatics 10330, pp. 396-402.

3. G. Tamazian, J. Ho Chang, **S. Knyazev**, E. Stepanov, K.J. Kim, Y. Porozov, "Modeling conformational redox-switch modulation of human succinic semialdehyde dehydrogenase," **Proteins**. 2015 Dec;83(12):2217-29.

2. **S.N. Knyazev**, V.Y. Kalyakin, I.N. Deryabin, B.A. Fedorov, A.V. Smirnov, E.O. Stepanov, and Yu.B. Porozov, "Prediction of protein conformational mobility with validation using small-angle X-ray scattering," **Biophysics** 60, 886–892 (2015)

1. **S.N. Knyazev**, V.Y. Kalyakin, I.N. Deryabin, B.A. Fedorov, A.V. Smirnov, E.O. Stepanov, and Yu.B. Porozov, "Prediction of protein conformational mobility with validation using small-angle X-ray scattering," **Biofizika** 2015 Nov-Dec;60(6):1069-1076

1.6 Refereed Articles in Conference Proceedings

7. A. Melnyk, F. Mohebbi, **S. Knyazev**, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovsky, M. Patterson, "Clustering Based Identification of SARS-CoV-2 Subtypes," In: Jha S.K., Măndoiu I., Rajasekaran S., Skums P., Zelikovsky A. (eds) Computational Advances in Bio and Medical Sciences. ICCABS 2020. Lecture Notes in Bioinofrmatics 12686, 127-141, 2021.

6. K.Mitchell, J.J.Brito, I.Mandric, Q.Wu, **S.Knyazev**, S.Chang, L.S.Martin, A.Karlsberg, E.Gerasimov, R.Littman, B.L.Hill, N.C.Wu, H.Yang, K.Hsieh, L.Chen, E.Littman, T.Shabani, G.Enik,

D.Yao, R.Sun, J.Schroeder, E.Eskin, A.Zelikovsky, P.Skums, M.Pop, S.Mangul: "Benchmarking of computational error-correction methods for next-generation sequencing data," BCB 2020: 63:1

5. F. Rondel, R. Hosseini, B. Sahoo, **S. Knyazev**, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, A. Zelikovsky, "Estimating Enzyme Participation in Metabolic Pathways for Microbial Communities from RNA-seq Data," Proc. of International Symposium on Bioinformatics Research Applications (ISBRA), 2020, Lecture Notes in Bioinformatics 12304, 335-343

4. I. Mandric, **S. Knyazev**, C. Padilla, F. Stewart, I. I. Măndoiu, and A. Zelikovsky, "Metabolic Analysis of Metatranscriptomic Data from Planktonic Communities," Proc. of International Symposium on Bioinformatics Research Applications (ISBRA), 2017, Lecture Notes in Bioinformatics 10330, pp. 396-402.

3. E. Lutsenko, **S. Knyazev**, Yu. Porozov (2016) A new approach for modeling of conformational changes of multi chain proteins. 5th International Young Scientists Conference in High Performance Computing and Simulation (YSC) Procedia Computer Science, DOI: 10.1016/j.procs.2016.11.009

2. **S. Knyazev**, S. Tarakanov, V. Kuznetsov, Y. Koucheryavy, E. Stepanov, Yu. Porozov (2014) Coarse-Grained Model of Protein Interaction for Bio-inspired Nano-communication. 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp.260-262. DOI:10.1109/ICUMT.2014.7002112

1. M. Buzdalov, **S. Knyazev** and Yu. Porozov (2014) Proceedings of Protein Conformation Motion Modeling using sep-CMA-ES. 2014 13th International Conference on Machine Learning and Applications (ICMLA), pp. 35-40. DOI:10.1109/ICMLA.2014.12

Invited Talks

7. P. Skums, A. Zelikovsky and **S. Knyazev** Computational analysis of viral genomes and outbreaks. 11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (Tutorial at ACM-BCB), 2020

6. **S. Knyazev** and A. Zelikovsky Reconstruction of Intra-Host Viral Populations Using Next Generation Sequencing. 9th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2019

5. **S. Knyazev** Reconstruction of Intra-Host Viral Populations Using Next Generation Sequencing Advanced molecular detection at Centers for Disease Control and Prevention, Research in Progress, Division of HIV Prevention, Centers for Disease Control and Prevention, 2019

4. **S. Knyazev** and Alex Zelikovsky CliqueSNV: Scalable Reconstruction of Intra-Host Viral Populations from NGS Reads The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB), 2018

3. **S. Knyazev**, V. Tsyvina, A. Artyomenko, P. Skums, A. Zelikovsky Accurate Reconstruction of Viral Population Using Correlation between SNVs. 7th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2017

2. **S. Knyazev** VOICE: Viral Outbreak Inference Tool. Annual Retreat, Molecular Basis of Disease Area of Focus, Georgia State University, 2017

1. O. Glebova, **S. Knyazev**, A. Artyomenko, A. Zelikovsky Simulation-based inference of genetic relatedness between viral populations. 6th IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2016

CHAPTER 2

EPIDEMIOLOGICAL DATA ANALYSIS OF VIRAL QUASISPECIES IN THE NEXT-GENERATION SEQUENCING ERA

2.1 Viral populations analysis problem and challenges

The NGS extracts quantitatively and qualitatively more information from patients' viral samples than the Sanger sequencing. But the extraction of this information requires sophisticated algorithms and software tools. In the following, we have reviewed bioinformatics methods and tools for NGS data analysis in viral epidemiology which can be partitioned into the following three categories (see Figure 1.1):

- Primary sequencing data analysis that consists of main strain reconstruction, read alignment and characterization of intra-host viral population structure including SNV and haplotype calling.

- Secondary sequencing data analysis that employs reconstructed viral populations for predicting drug resistance, estimating recency of infection, and outbreak investigation, including transmission cluster detection and identification of transmission direction and outbreak sources.

- Molecular surveillance systems that provide a software environment for combined primary and secondary analysis of viral NGS data in real-time.

NGS-based characterization of intra-host viral population structures is advanced enough and is getting ready to be used in epidemiological and clinical studies. This claim is supported by the

number of recently published studies that use quasispecies analysis for outbreak investigation and transmission inference[142,3,140]. Inferred intra-host viral population structure can facilitate accurate answers to essential epidemiological questions about drug-resistance, recency of infection, transmission clusters and outbreak sources. Future NGS-based surveillance systems should employ big data analytics to combine enormous amounts of sequencing and epidemiological data for the timely detection of outbreaks and the design of efficient public health intervention strategies.

2.2 The primary analysis of viral next-generation sequencing data

Primary analysis can be partitioned into two major steps: (i) basic primary analysis which starts with error correction followed by identification of the consensus sequence and read mapping and (ii) characterization of the intra-host viral population complexity by calling single nucleotide variants (SNV) and haplotype variants in the viral sample.

2.3 Basic primary analysis

The error correction of viral sequencing reads is a notoriously difficult task. The standard error correction tools tuned to correct reads from a human genome do not perform well for viral genomes since viral haplotypes differ only slightly between themselves[120]. There are several error-correction tools that have been proposed specifically to handle viral sequencing samples[188,156,110]. A Bayesian probabilistic clustering approach[188] integrates error correction with SNV and haplotype calling, while KEC[156] is a k-mer counting-based approach that identifies erroneous k-mers by analyzing the distributions of k-mer frequencies. A more sophisticated random forest classifier MultiRes[110] can be used to distinguish between erroneous and rare k-mers.

Identification of the consensus sequence can be either picked from existing reference genomes or *de novo* assembled in order to avoid reference biases. The reference-based identification of the consensus relies on the existence of closely related genomic sequences. NGS reads are aligned to the reference sequence with a significant number of mismatches. In order to avoid reference biases, the aligned reads are used for updating each position of the reference genome with the base most frequent in reads and re-aligning reads to the consensus[12,80]. The drawback of this approach is that selecting the reference genome is not a well-formalized procedure.

De novo assemblers are based on De Bruijn graphs such as VICUNA and overlap graphs such as SAVAGE[76,175,185,83,13]. SAVAGE constructs an overlap graph with vertices representing reads and/or contigs and edges connecting two reads/contigs belonging to the same haplotypic sequence. Statistically, well-calibrated groups of reads/contigs are then efficiently used for reconstruction of the individual haplotypes from this overlap graph. SAVAGE has an additional advantage over VICUNA since it builds multiple haplotype contigs rather than a single consensus. *De novo* assemblers require much higher memory and time resources than reference-based identification of the consensus.

A recent tool, SHIVER[183], combines the reference-based and *de novo* approaches by using both reads and contigs assembled from those reads for HIV sequencing. Contigs are compared with the existing references, wherein some are spliced and some are removed as contaminants. After the closest existing reference is identified it is updated to the consensus by well-mapped reads that do not match contaminants.

2.3.1 Single nucleotide variant calling

The natural advantage of NGS vs Sanger sequencing is its ability to identify low-frequency mutations (i.e., below 20%) that are particularly relevant in the context of drug resistance[19,8,167]. The main challenge for SNV calling is to distinguish between sequencing errors and low-frequency true SNVs. All existing methods apply a particular error model to estimate the probability that an observed mismatch with the consensus is an error and qualify it as an SNV if this probability is low enough.

Below we briefly describe widely known tools (see[132]) and recently developed tools. VarScan[96] reports SNVs which are deeply covered by the reads with high quality. A similar approach with improved codon-based filtration is introduced by VirVarSeq[170] of SNV. The method LoFreq[181] derives sequencing error probability from a Phred-scaled quality value and optimizes estimation of P-value. V-Phaser[108] introduces a basic primary analysis and error model, which takes into account the simultaneous occurrence of pairs of SNV in the same reads. V-Phaser 2[186] specifies this model for Illumina reads. Pairs of mutations are explored by CoVaMa[149] using a linkage disequilibrium model. An accurate analysis of linked SNV pairs independent of error rate is proposed by CliqueSNV[95] which also contains an efficient implementation of the SNV-pair analysis. ViVan[85] and ViVaMBC[169] are based on maximum likelihood models. MinVar[82] and SiNPle[57] utilize the Poisson–Binomial distribution and Bayesian model respectively. Validation of MinVar on Illumina Miseq samples and shows that SNVs with the frequency of at least 5% are reliably identified without introducing false-positives. PASeq[125] and Hydra Web[86] are web-based publicly available tools that are thoroughly tested for identifying mutations with frequencies 20% and 5%. Interestingly,

SNV calling for viral data is very similar to somatic mutation calling and the quality of algorithms for both problems can be compared[57].

Table 2.1 describes the list of tools analyzing viral NGS data for SNV calling. For each tool, we specify the SNV detection method and whether it requires a reference.

Table 2.1: SNV calling software tools for viral NGS data

SNV calling tools	Year	System	De-novo/ ref-based	Pair-end reads	SNV detection method	Tool availability
VarScan	2009	Java	ref	+	Read coverage	http://varscan.sourceforge.net/
LoFreq	2012	Linux	ref	+	Poisson–binomial distribution	https://csb5.github.io/lofreq/
Vphaser	2012	Linux	ref	-	Bernoulli phasing model	https://www.broadinstitute.org/viral-genomics/v-phaser
Vphaser2	2013	Linux	ref	+	Bernoulli phasing model	https://www.broadinstitute.org/viral-genomics/v-phaser-2
ViVan	2015	-	ref	+	Maximum likelihood	http://www.vivanbioinfo.org
ViVaMBC	2015	R	ref	+	Maximum likelihood	https://sourceforge.net/projects/vivambc/
VirVarSeq	2015	Linux	ref	+	Codon-level quality filtration	https://sourceforge.net/projects/virtools/?source=directory
CoVaMa	2015	Python	ref	+	Linkage disequilibrium	https://sourceforge.net/projects/covama/
MinVar	2017	Python	ref	+	Poisson–binomial distribution	http://git.io/minvar
MultiRes	2017	Linux	de-novo	+	Frame-based model	https://github.com/raunaq-m/MultiRes
CliqueSNV	2018	Java	ref	+	Linkage of SNV pairs	https://github.com/vtsyvina/CliqueSNV
SiNPle	2019	Linux	ref	+	Bayesian model	https://mallorn.pirbright.ac.uk:4443/gitlab/drcyber/SiNPle
PASeq		web				https://paseq.org/
Hydra Web		web				https://hydra.canada.ca/pages/home?lang=en-CA
SmartGen		web				https://www.smartgene.com/mod_hiv.html

2.3.2 Viral haplotype variant calling

Rather than determining variation in a single position, the haplotype calling is required to find the haplotypes spanning the entire viral genome or amplicons of special interest. The haplotypes and their frequencies are more informative than SNVs for detecting drug resistance which can non-linearly depend on accumulated SNVs. Haplotypes are also used for significantly more accurate detection of transmission clusters and outbreak sources.

Note that haplotype frequency reconstruction is considered to be a simpler problem as soon as haplotypes are inferred. The expectation-maximization algorithm based on the estimation of the probability that a given read has been emitted by a given haplotype has been shown to be sufficiently reliable with accuracy growing with the sequencing depth[12,189].

The first haplotype reconstruction tools were read-graph based with vertices corresponding to reference-mapped reads and edges connecting reads that agree on their overlap[55,180]. Many tools followed this idea[12,189,112,81,134,157,165,114,87,37] significantly improving the quality of reconstruction (see[132,113]). But all these tools usually are not fast enough to handle recently available multi-million read data sets.

Probabilistic modeling of the sequencing process and/or viral haplotype generation[89,164,133,103,109] was shown to be an attractive alternative to the read-graph approach. The most successful tool among probabilistic tools is PredictHaplo[133] that exhibits high specificity and can reconstruct haplotypes with frequency over 10%. Hierarchical-clustering of reads (especially long PacBio reads) has been suggested in[9], and recent methods aBayesQR[2] combined probabilistic modeling with clustering making the Bayesian approach computationally tractable.

Novel scalable tools handling millions of reads and improving over existing tools are actively developed in multiple labs. CliqueSNV[95] efficiently recognizes groups of linked SNVs and constructs an SNV graph, where SNVs are nodes and edges connect linked SNVs It can assemble close viral haplotypes with frequencies as low as 0.1% from Illumina and PacBio reads.

It is necessary to separately note de novo haplotype callers, i.e., tools that de novo assemble multiple distinct haplotypes rather than a consensus. Currently, there exist three de novo assemblers MLEHaplo[109], SAVAGE[13] and PEHaplo[37]. The advantage of these tools is that they do not introduce reference biases.

Recently, twelve NGS haplotype callers were tested using viral populations simulated under realistic evolutionary dynamics but without error simulation[54]. In contrast to other simulations, the number of haplotypes was very large (216 -1,185) and each frequency was small ($< 7\%$). Under such stressful conditions, PreditHaplo and CliqueSNV showed certain advantages over other reference-based methods and PEHaplo among de novo assemblers.

Table 2.2 describes the list of tools analyzing viral NGS data for haplotype calling. For each tool, we specify (1) whether it is a de novo method or requires a reference, (2) sequencing error handling, (3) the method for haplotype assembly, (4) and the method for haplotype frequency estimation.

Table 2.2: Haplotype calling software tools for viral NGS data

Haplotyping tools	**Year**	**System**	**De-novo / ref-based**	**Pair-end reads**	**Sequencing error handling**	**Haplotype assembly method**	**Haplotype frequency estimation method**	**Output sequences**	**Tool availability**
Shorah	2011	Linux	ref	+	Probabilistic clustering	Minimal path cover	EM	Full haplotypes	https://github.com/cbg-ethz/shorah
ViSpA	2011	Linux	ref	-	Binomial model	Max-bandwidth path	EM	Full haplotypes	http://alan.cs.gsu.edu/NGS/?q=content/vispa
QColors	2012	-	de-novo	-	-	Overlap graph + Conflict graph	-	Full haplotypes	-
QuRe	2012	Java	ref	+	Poison model	Multinomial distribution matching	Read coverage	Full haplotypes	https://sourceforge.net/projects/qure/
bioa	2012	Linux	ref	-	k-mer based error correction	Maximum Bandwidth Path	Fork balancing	Full haplotypes	http://alan.cs.gsu.edu/vira/index.html
Vicuna	2012	Linux	de-novo	+	Read count	-	-	consensus + contigs	https://www.broadinstitute.org/viral-genomics/vicuna
QuasiRecomb	2013	Linux	ref	+	Hidden Markov model	Hidden Markov model	Hidden Markov model	Full haplotypes	https://github.com/cbg-ethz/QuasiRecomb
Vira (AmpMCF)	2013	Linux	ref	-	-	Multicommodity Flows	Normalized flow size	Full haplotypes	http://alan.cs.gsu.edu/vira/index.html
ShotMCF	2013	JAVA	ref	-	Binomial model	Max-bandwidth path + Multicommodity Flows	EM + Normalized flow size	Full haplotypes	http://alan.cs.gsu.edu/NGS/?q=content/shotmcf
BAsE-Seq	2014	-	ref	+	Poisson–binomial distribution model	Clustering of reads by SNVs	Read coverage	Full haplotypes	-
VGA	2014	Linux	ref	+	Requires high-fidelity sequencing protocol	Min-graph coloring	EM	Full haplotypes	http://genetics.cs.ucla.edu/vga/
HaploClique	2014	Linux	ref	+	-	Max-clique enumeration	Normalized read count	Full haplotypes	https://github.com/cbg-ethz/haploclique
PredictHaplo	2014	Linux	ref	+	Dirichlet Process Mixture Model	Dirichlet Process Mixture Model	Dirichlet Process Mixture Model	Full haplotypes	https://bmda.dmi.unibas.ch/software.html

continued on next page

continued from previous page

IVA	2015	Linux	de-novo	-	Read count	-	-	contigs	https://sanger-pathogens.github.io/iva/
MLEHaplo	2015	Linux	de-novo	+	-	Maximum Likelihood	-	Full haplotypes	https://github.com/raunaq-m/MLEHaplo
ViQuaS	2015	Linux	ref	+	Chimeric error correction	Multinomial distribution matching	Read count	Full haplotypes	https://sourceforge.net/projects/viquas/
SAVAGE	2017	Linux	de-novo	+	Overlap fuzzy matching error correction	Enumerating cliques in overlap graph	EM	contigs	https://bitbucket.org/jbaaijens/savage/
aBayesQR	2017	Linux	ref	+	Cluster coverage by reads	Bayesian inference	Bayesian inference	Full haplotypes	https://github.com/SoYeonA/aBayesQR
RegressHaplo	2017	R	ref	+	-	Penalized Regression	Penalized Regression	Full haplotypes	https://github.com/SLeviyang/RegressHaplo
2SNV	2017	Java	ref	-	Linkage of SNV pairs	Hierarchical clustering of reads by SNVs	EM	Full haplotypes	http://alan.cs.gsu.edu/NGS/?q=content/2snv
PEHaplo	2018	Linux	de-novo	+	Overlap error correction	Path finding in overlap graph	-	contigs	https://github.com/chjiao/PEHaplo
Shiver	2018	Linux	de-novo + ref	+	BLAST database match	-	-	consensus	https://github.com/ChrisHIV/shiver
CliqueSNV	2018	JAVA	ref	+	Linkage of SNV pairs	Clique enumeration and merging	EM	Full haplotypes	https://github.com/vtsyvina/CliqueSNV

2.4 Secondary analysis of viral next-generation sequencing data

Secondary NGS analysis addresses three tasks: (i) predicting of drug resistance which takes SNV and haplotypes obtained during primary analysis and determine whether they are drug-resistant or not; (ii) determining the recency of the infection, that is predicting the moment in the past when patient was infected; (iii) outbreak investigation, that is determining the borders of outbreak, finding the source of infection, and reconstruction of infection spread paths.

2.4.1 Predicting drug resistance

Certain haplotypes and mutations that are found during the primary NGS should be analyzed for drug resistance. This is especially important for viruses such as HIV[104], HCV[148], influenza[130], and others[84]. For HIV, the detection of drug resistance is especially relevant since HIV patients have to adhere to a treatment for the span of their lives. If a patient develops HIV drug resistance they will be required to switch to a different line of treatment, and these treatments may be less studied and of a higher risk to the patient's health. Additionally, the number of drug-resistant mutations in the patient is constantly growing as well as the number of drug-resistant patients in the outbreak[68]. This makes the task of tracking HIV drug resistance a more onerous one[10].

Detection of drug resistance is typically associated with matching genome mutations with the efficiency of drugs[84]. Usually, different mutations have different resistance power and often mutations work collectively[62], so the process of finding correlations between mutations and drug resistance is non-linear[56]. The comprehensive overview of computational approaches to drug-resistant HIV mutations can be found in[145]. Most of the tools are aimed at Sanger sequencing data since

NGS data has only been accumulating for a short period of time. Sanger sequencing allows the detection of mutations with frequencies above 25% which has low benefits for the clinical application[100,49]. NGS increases the sensitivity and lowers the frequency threshold up to 1-5%[74].

There are two main challenges in the detection of drug resistance that depends on the results of primary NGS data analysis. They are connected with the accuracy of detecting minority mutations and haplotypes. The first problem is that if there is a minor drug-resistant mutation, the haplotypes with this mutation will have an advantage over other haplotypes dealing with drug pressure. As a result, these drug-resistant haplotypes will begin to dominate over time[104,88]. The second problem is that drug resistance is connected with haplotypes rather than with the mutations themselves, but haplotypes are harder to detect and so the drug resistance analysis can be significantly improved with more sensitive haplotyping tools[128].

Currently, tools for detecting drug resistance are modeled to handle Sanger sequencing data accumulated in designated databases[145]. The limitation of Sanger data is that only the major haplotype and SNVs with frequency at least 20% can be reconstructed. This hurts the performance of the most efficient drug resistance prediction tools that are based on machine-learning[64,128,126,182,17,153]. Such tools would rather take into account all patient's haplotypes[128,35] to overcome Sanger sequencing limitations by generating all possible haplotypes with given SNVs, e.g., 10 SNVs make $2^{10} = 1024$ different haplotypes.

The number of HIV patients sequenced with NGS is beginning to grow very fast. Since NGS can detect rare SNPs and haplotypes, drug resistance can be predicted more accurately[62,145]. We expect that the number of NGS samples to train these models will grow much faster after the

FDA authorizes the first next-generation sequencing test for detecting HIV-1 drug resistance mutations[143]. Recent clinical studies showed up to 2.7-fold improvement for detecting drug resistance with utilizing NGS data[119,60,4,63,167,43,42] to antiretroviral therapy such as Zidovudine (see Table 2.3).

Table 2.3 Detection of drug-resistant mutations in clinical studies: NGS vs Sanger sequencing

Study	**Patients group**	**Patients number**	**Collection date**	**Region**	**DRM detection: NGS/Sanger (fold)**
Metzner et al. 2005	acute patients	49	1999-2003	Germany	2.0
Fisher et al. 2015	infants after PMTCT failure	15	2006-2009	South Africa	2.5
Alidjinou et al. 2017	ART-naive patients	48	2013-2015	France	2.7
Tzou et al. 2018	Undisclosed	177	2001-2016	Undisclosed	1.2
Fokam et al. 2018	Vertically infected children	18	2015	Cameroon	1.7
Derache et al. 2019	ART-naive patients	1148	2012-2016	South Africa	1.4
Derache et al. 2019	Patients failing 1st line ART	1287	2012-2016	South Africa	2.0

2.4.2 Estimating infection recency

Over 80% of untreated cases of HCV infection becomes chronic. This impedes the timely diagnosis of the disease, due to the fact that the infection often does not manifest any clinical symptoms in its early stages. Currently, there are no diagnostic assays to determine the stage of HCV infection. Therefore, distinguishing recently infected patients from chronically infected patients using computational methods would be highly advantageous for both personalized therapeutic purposes and for epidemiological surveillance; e.g., for detection of incident HCV cases. Similarly, detection of the age of HIV infection is crucial for HIV-1 surveillance and the understanding of viral pathogenesis[34].

Measuring the time since infection using genomic data has recently been addressed in several studies[34,122,11,16,15]. The simpler version of this problem is infection staging, i.e. distinguishing between recent and chronic infections using viral sequences sampled by NGS. A number of methods establish an age or stage of HIV or HCV infection using various measures of the population structure[34,122,11,16,15]. An underlying assumption of such methods is that intra-host viral evolution is associated with continuous genetic diversification. This results in the existence of a correlation between genetic heterogeneity of quasispecies and the age of quasispecies, which allows for the use of properly calibrated diversity measures as age markers.

Recently, groups of comprehensive features accounting for population diversity, population genetics, topological, information-theoretical and physico-chemical properties of quasispecies populations were integrated using sophisticated machine-learning-based techniques[16,15]. These methods take into account recent observations in the evolution of viruses, such as HCV, resulting in a

gradual intra-host adaptation that is accompanied by a decrease in heterogeneity and an increase in negative selection[27,139,69,48].

2.4.3 Outbreak investigation

Detection and investigation of viral outbreaks is the primary epidemiological task. Historically, epidemiological investigations have been based on in-field surveys of epidemiological settings and interviews with persons potentially involved in pathogen spread. However, such methods are time- and labor-consuming and the data obtained is prone to various socio-behavioral biases. Analysis of viral genomic data provides alternative unbiased machinery for outbreak investigations and quantification of major factors responsible for disease spread[127].

It should be noted that in the recent decade the rich variety of tools for inferring epidemiological parameters has been developed within the field of viral phylodynamics[141,171]. In addition, there are a plethora of methods for outbreak investigations that combine various types of genomic and epidemiological data[171,94,90,40,91,121,123,187]. Despite being highly effective in many settings, these tools are currently not intended for application to NGS data and usually do not support calculations with extremely large genomic datasets. Therefore in this paper, we concentrate on tools specifically designed to handle heterogeneous intra-host viral populations using NGS.

The primary task in the outbreak investigation is the detection of transmission clusters. The main challenge here is the development and implementation of evolutionary distance measures between intra-host viral populations that reflect the epidemiological relations between the hosts. These distances can be efficiently calculated and combined with a broad variety of clustering techniques and phylogenetic and network-based methods[26,5]. Distances between consensus sequences

that are still often used for epidemiological investigations provide only very coarse estimates of evolutionary distances and lose significant signal encoded in quasispecies structure. In particular outbreak distances between viral variants from certain hosts can be comparable or even higher than distances between variants from different hosts. For example, for HIV-1, the recommended inter-host threshold for detecting transmission clusters in pol region is in a range of 0.5 - 1.5%[127], while the nucleotide genetic variability inside hosts can be as high as 5%[152].

Analysis of quasispecies populations reconstructed from NGS data drastically improves the estimation of evolutionary distances. Pioneering NGS-based study for HCV outbreak investigations[30] proposed to measure the distance between samples as the distance between the closest pair of haplotypes from different samples. Even this simple method has been shown to significantly outperform the consensus-based approach[30]. Similar techniques have been applied to HIV[97]. Despite the simplicity of the metric, its calculation is challenging for extremely large NGS datasets, since its naive implementation requires a pairwise comparison of sequences from all pairs of patients. To address this challenge, several filtering techniques have been proposed[151,166]. In consecutive studies[160,70,15,118] more sophisticated distance measures for quasispecies populations have been proposed. In particular,[118] avoids reconstruction of haplotypes and/or phylogenetic trees by utilizing k-mer-based approach. Specifically, each viral sample is represented by a corresponding k-mer distribution, the distance between pairs of k-mers is computed over a single de Bruijn graph of all k-mers, and the distance between populations is identified with the Earth Mover's Distance (EMD) between two k-mer distributions.

The next step of the bioinformatics pipeline for epidemiological analysis is an investigation

of viral transmissions inside each transmission cluster. It includes a prediction of possible transmission directions, detection of the source or "superspreader" of an outbreak and inference of transmission networks indicating who infected whom. QUENTIN[160] and VOICE[70] estimate the distance between quasispecies populations as the analog of a cover for a Markov-type model of viral evolution and choose the direction of transmission from a sample A to sample B based on minimum evolution principle, i.e. if it requires less evolution time than the time for evolving from A to B. In[147], it is proposed to identify the transmission directions by phylogenetic analysis and detection of paraphyletic, polyphyletic and monophyletic relations between sampled intra-host variants from different hosts. This idea has been further developed and implemented in Phyloscanner[184].

Both QUENTIN and Phyloscanner also allow reconstructing viral transmission networks. QUENTIN does it via Bayesian inference and Markov Chain Monte Carlo sampling, with the likelihood of a transmission network being defined using general properties of social networks relevant to the infection dissemination. Phyloscanner relies on a maximum-parsimony approach and assigns ancestral hosts to internal nodes of a viral phylogeny containing quasispecies populations from different hosts by minimizing the number of transmission events while taking into account possible contaminations, multiple infections, and presence of unsampled hosts.

Before determining the source of the outbreak it is critical to decide whether the source of the outbreak is present among sequenced samples[118]. Finding the source of an outbreak is quite important for outbreak disruption. The papers[160,70,118] validated their approaches on CDC data for HCV outbreaks with the known sources and showed that the source prediction accuracy is around 90%. But before determining the source of the outbreak it is critical to decide whether the source

of the outbreak is present among sequenced samples[118]. This problem is quite difficult and has been addressed for the first time in[118].

Table 2.4 describes the list of tools analyzing viral NGS data for outbreak investigation including identification of (1) transmission clusters, (2) transmission direction, (3) source of infection, (4) presence of source, (5) transmission network. For each tool we indicate which of five tasks are addressed by which tool.

Table 2.4 Outbreak investigation software tools for viral NGS data

Tool	Year	System	Algorithm	Transmission clusters	Transmission direction	Transmission network	Source of infection	Presence of source	Tool availability
MinDist	2016	-	Distance based	+	-	-	+	-	-
RED	2017	Matlab	Clustering	+	+	-	+	-	https://bitbucket.org/osaofgsu/red
VOICE	2017	Linux	Simulation based	+	+	-	+	-	https://bitbucket.org/osaofgsu/voicerep
PhyloScanner	2017	Linux	Phylogeny	+	+	+	+	-	https://github.com/BDI-pathogens/phyloscanner
Quentin	2017	Matlab	Simulation based	+	+	+	+	-	https://github.com/skumsp/QUENTIN
signature-sj	2018	Java	k-mers	+	-	-	-	-	https://github.com/vtsyvina/signature-sj
k-mer EMD	2019	Linux	k-mer based distance	+	+	-	+	+	https://github.com/amelnyk34/kemd

2.5 Molecular surveillance systems and databases

The advent of next-generation sequencing technologies makes possible, for the first time, the deployment of molecular epidemiological surveillance systems that are intended to analyze and infer the dynamics of epidemics and outbreaks in real or almost real-time using computational analysis of viral genomic data[97,105]. Such systems are characterized by a broad bioinformatics functionality including the processing of raw sequencing data, sequence alignment, phylogeny or network construction, transmission history inference and visualization. The number of computational molecular surveillance systems are currently being developed and deployed. One of the widely cited systems is Nextstrain[72] that allows for phylodynamics analysis and interactive visualization of the evolution of a variety of pathogens. The Nextstrain incorporates several computational tools for alignment, phylogenetic inference, reconstruction, dating and geographic localization of transmission events. However, currently, a toolkit of Nextstrain is not intended for the analysis of next-generation sequencing data and intra-host viral populations, although its open-source architecture makes possible incorporation of such methods in the future. The library of tools for viral epidemiological data analysis developed and maintained by the R Epidemics Consortium (RECON) also should be mentioned. It includes R statistical packages for handling, visualizing, and analyzing outbreak data, but has similar limitations.

Two surveillance systems that support NGS data are specifically tailored for HIV and Viral Hepatitis and are recommended and/or maintained by the CDC. These systems are HIV-Trace[97] and GHOST (Global Hepatitis Outbreak Surveillance Technology)[105], and they are based on high-throughput bioinformatics pipelines for genetic relatedness analysis. They allow estimates of ge-

netic distances between intra-host populations sampled from HIV-infected individuals, use these distances to detect possible transmission linkages between the individuals, reconstruct and visualize transmission clusters and genetic relatedness networks. Both systems can work with haplotypes obtained from NGS data and are scalable for extremely large datasets produced by Illumina MiSeq and other sequencing platforms. In particular, GHOST employs several efficient k-mer-based filtering techniques for viral sequence similarity queries, that allow for the elimination of an exhaustive comparison of all pairs of viral haplotypes and allow processing of NGS data from a given HCV outbreak in minutes[166].

Another important issue is the creation of curated databases that contain both genomic and epidemiological data and can be used for the validation of new computational molecular epidemiology tools. Some previously published papers[160,70] provide links to datasets that can be used for these purposes, but, to the best of our knowledge, large systematically curated collections of such datasets are yet to be created. In this context, Pangea HIV consortium efforts on curated analysis for HIV outbreaks in the African region[1] are very important. At this moment they maintain a collection of more than 18000 HIV NGS samples that can be used for outbreak investigations and data-driven design of prevention strategies.

CHAPTER 3

CLIQUESNV - A METHOD FOR INFERRING VIRAL QUASISPECIES USING NGS

Background

Rapidly evolving RNA viruses such as influenza A virus (IAV), human immunodeficiency virus (HIV) and hepatitis C virus (HCV) exist in infected hosts as highly heterogeneous populations of closely related genomic variants called quasispecies[92,73,107,53,115,162,47,146].

The composition and structure of intra-host viral populations plays a crucial role in disease progression and epidemic spread. The presence of low-frequency variants that differ from major strains by a few mutations may result in immune escape, emergence of drug resistance, and an increase of virulence and infectivity[18,50,65,79,144,29,158]. Furthermore, such minor variants are often responsible for transmissions and establishment of infection in new hosts[31,59,6]. Therefore, accurate characterization of viral mutation profiles sampled from infected individuals is essential for viral research, therapeutics and epidemiological investigations.

Next-generation sequencing (NGS) technologies now provide versatile opportunities to study viral populations. In particular, the popular Illumina MiSeq/HiSeq platforms produce 25-320 million reads, which allow multiple coverage of highly variable viral genomic regions. This high coverage is essential for capturing rare variants. However, *haplotyping* of heterogeneous viral populations (i.e., assembly of full-length genomic variants and estimation of their frequencies) is extremely complicated due to the vast number of sequencing reads, the need to assemble an unknown number of closely related viral sequences and to identify and preserve low-frequency variants. Single-molecule sequencing technologies, such as PacBio, provide an alternative to

short-read sequencing by allowing full-length viral variants to be sequenced in a single pass. However, the high level of sequence noise (due to background or platform-specific sequencing errors) produced by all currently available platforms makes inference of low-frequency genetically close variants especially challenging, since it is required to distinguish between real and artificial genetic heterogeneity produced by sequencing errors.

In recent years, a number of computational tools for inference of viral quasispecies populations from noisy NGS data have been proposed, including Savage[13], PredictHaplo[133], aBayesQR[2], QuasiRecomb[164], HaploClique[165], VGA[114], VirA[157,112], SHORAH[189], ViSpA[11], QURE[134] and others[159,14,180]. Even though these algorithms proved useful in many applications, accurate and scalable viral haplotyping remains a challenge. In particular, inference of low-frequency viral variants is still problematic, while many computational tools designed for the previous generation of sequencing platforms have severe scalability problems when applied to datasets produced by state-of-the-art technologies.

Previously, several tools such as V-phaser[108], V-phaser2[186] and CoVaMa[149] exploit linkage of mutations for single nucleotide variant (SNV) calling (rather than haplotype assembly), but they do not take into account sequencing errors when deciding whether two variants are linked. These tools are unable to detect mutations of frequency above sequencing error rates[170]. The 2SNV algorithm[9] accommodates errors in links and was the first such tool to be able to correctly detect haplotypes with a frequency below the sequencing error rate.

Other methods (e.g., HaploClique[165], Savage[13]) assembled viral haplotypes using maximal cliques in a graph, where nodes represent reads. These methods infer haplotypes by iteratively

merging these read cliques, thus heavily relying on the correct order of merging. In contrast, our proposed approach finds maximal cliques in a graph with nodes corresponding to SNVs, which facilitates a significant performance increase, since for viruses the size of the SNV graph is significantly smaller than the size of the read graph. Furthermore, the clique merging problem is formulated and solved as a combinatorial problem on the auxiliary graph of cliques of the SNV graph, thus allowing an increase of the CliqueSNV algorithm's accuracy.

Herein, we present CliqueSNV, a novel method that is designed to infer closely related intra-host viral variants from noisy next-generation and third-generation sequencing technologies[22]. It infers haplotypes from patterns from distributions of SNVs inside sequencing reads. CliqueSNV is suitable for long single-molecule reads (PacBio) as well as for short paired-end reads (Illumina). Our method recognizes groups of linked SNVs and efficiently distinguishes them from sequencing errors. CliqueSNV constructs an SNV graph, where SNVs are nodes and edges connect linked SNVs. Then, by merging cliques in that graph, CliqueSNV identifies true viral variants. Using optimized combinatorial techniques makes CliqueSNV fast and efficient in comparison with other tools.

Validation testing shows that CliqueSNV outperforms PredictHaplo[133], aBayesQR[2] and 2SNV[9] in both speed and accuracy using four real and two simulated datasets. Other haplotyping methods have been shown to perform similarly or worse than these three methods. Our benchmarks consist of sequencing experiments from samples with known viral mixtures: (i) a real PacBio sequencing experiment from a sample with ten influenza A (IAV) viral variants[9], (ii) two real MiSeq sequencing experiments from two samples of HIV-1 mixtures with nine and two viral variants, (iii) real

MiSeq data from a sample with five HIV-1 variants of different subtypes, and (iv) two simulated MiSeq datasets with IAV and HIV-1 sequences.

Together with standard precision and recall metrics we introduced two additional quality measures: (i) Matching Error between Populations and (ii) Earth Mover's Distance between Populations. These two measures are more adapted for evaluating the quality of inference of viral samples from sequencing data because they take into account both the distance between true and inferred haplotypes and the frequencies of true and inferred haplotypes.

3.1 CliqueSNV algorithm

The schematic diagram of the CliqueSNV algorithm is shown in Figure 3.1. The algorithm takes aligned reads as input and infers haplotype sequences with their frequencies as output. The method consists of six steps. Step 1 uses aligned reads to build the consensus sequence and identifies all SNVs. Then all pairs of SNVs are tested for dependency and are then divided into three groups: *linked*, *forbidden*, or *unclassified*. Each SNV is represented as a pair (p, n) of its position p and nucleotide value n in the aligned reads. If there are enough reads that have two SNVs (p, n) and (p', n') simultaneously, then they are tested for dependency. If the dependency test is positive and statistically significant (see Detailed description for details), then the algorithm classifies these two SNVs as *linked*. Otherwise, these two SNVs are tested for independency. If the independency test is positive and statistically significant (see Detailed description for details), then these two SNVs are classified as a *forbidden* pair. In Step 2, we build a graph $G = (V, E)$ with a set of nodes V representing SNVs, and a set of edges E connecting linked SNV pairs. Step 3 finds all *maximal*

cliques in graph G. A clique $C \subseteq V$ is a set of nodes such that $(u, v) \in E$ for any $u, v \in C$ and for any $x \notin C$ there is $u \in C$ such that $(x, u) \notin E$. Each maximal clique in G represents groups of pairwise-linked SNVs that potentially belong to a single haplotype. Ideally, there is a one-to-one correspondence between SNV cliques and haplotypes. Unfortunately, sequencing noise and the shortness of the NGS reads makes it difficult to identify all linked SNV pairs. As a result, a single clique corresponding to a haplotype will be split into several overlapping cliques. Step 4 merges such overlapping cliques. In order to avoid merging distinct haplotypes, two cliques are not merged if they contain a forbidden SNV pair. Step 5 assigns each read to a merged clique with which it shares the largest number of SNVs. Then CliqueSNV builds a consensus haplotype from all reads assigned to a single merged clique. Finally, haplotype frequencies are estimated via an expectation-maximization algorithm in Step 6.

Below we describe the six major steps of CliqueSNV that are schematically presented in Figure 3.1.

Step 1: Finding linked and forbidden SNV pairs. At a given genomic position I, the most frequent nucleotide is referred to as a *major variant* and is denoted 1. Let us fix one of the less frequent nucleotide (referred to as a *minor variant*) and denote it 2. A pair of variants at two distinct genomic positions I and J is referred to as a 2-haplotype. Let O_{22} be the observed count of the 2-haplotype (22) in the reads covering positions I and J. In this step, CliqueSNV tries to decide whether the observed O_{22} reads are sequencing errors or they are produced by an existing haplotype containing the 2-haplotype (22).

The pairs of minor variants (referred to as SNV pairs) are classified into three categories:

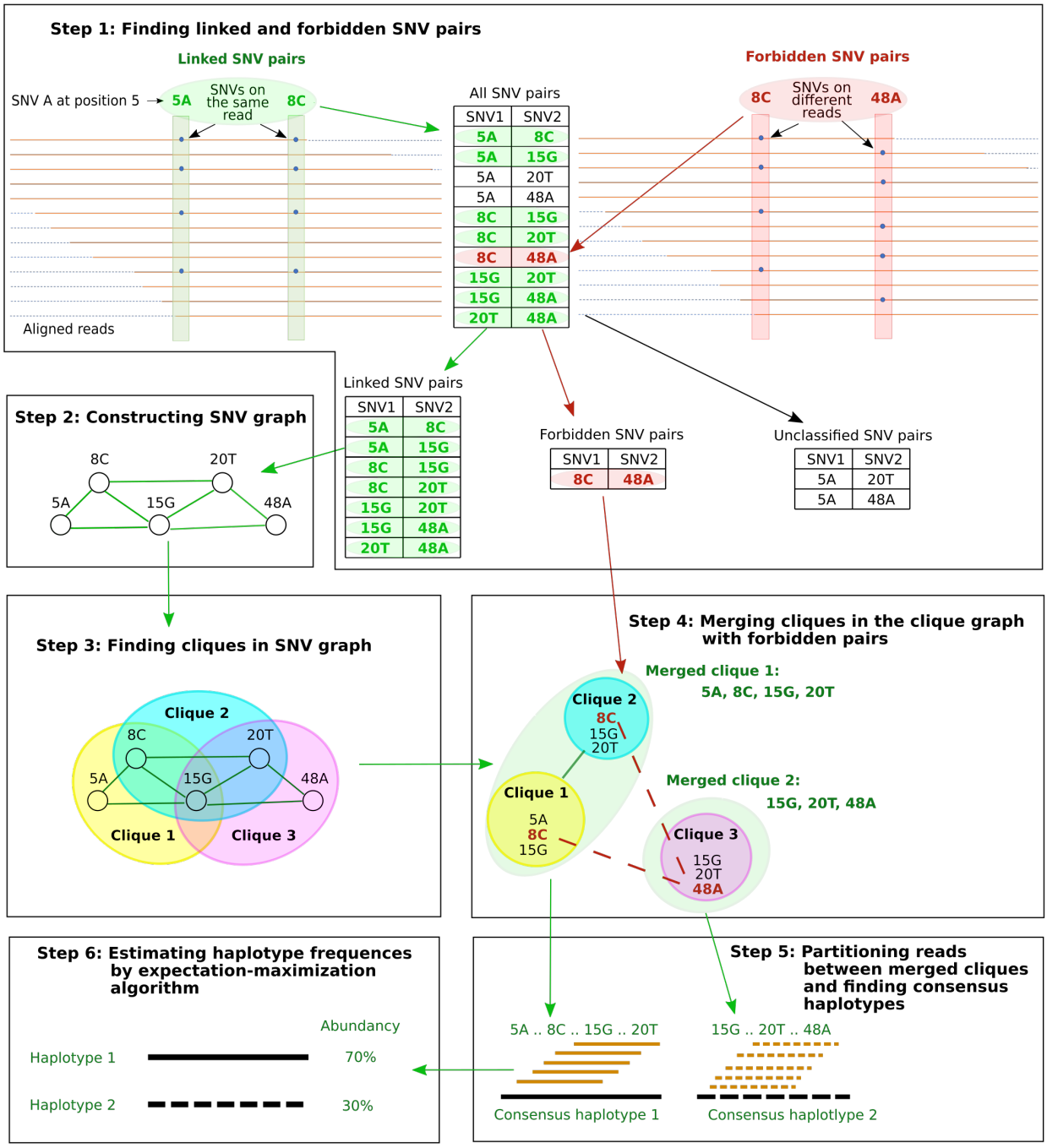

Figure 3.1 Schematic representation of the CliqueSNV algorithm, where SNV is single nucleotide variation.

linked, forbidden, and unclassified. An SNV pair is *linked* if it is extremely unlikely that there is no sufficiently frequent haplotype containing both minor variants is very low. On the other side, an SNV pair is *forbidden* if it is extremely unlikely that the corresponding minor variants belong to the same haplotype of sufficient frequency. All other SNV pairs are referred to as *unclassified*.

Below we estimate the probability of observing at least $x \geq O_{22}$ reads given that the true frequency T_{22} of the 2-haplotype (22) is at most t (by default $t = 0.001$). This probability should be low enough so that false positive linked pairs would be virtually impossible, i.e., we require that the expected number of false positive linked pairs be less than 0.05. Therefore, this probability should be less than $0.05/\binom{L}{2}$, where L is the haplotype length.

$$\begin{aligned} Pr(x \geq O_{22} | T_{22} \leq t) &= 1 - Pr(x < O_{22} | T_{22} \leq t) \\ &\leq 1 - \sum_{i=0}^{O_{22}-1} \binom{n}{i} t^i (1-t)^{n-i} \\ &\leq \frac{0.05}{\binom{L}{2}} \end{aligned} \tag{3.1}$$

Pairs of SNVs passing this linkage test (3.1) are classified as a *linked* SNV pairs.

For every other pair of SNVs, we check whether they can be classified as a *forbidden* SNV pair, i.e., whether the probability of observing at most $x \leq O_{22}$ reads is low enough (< 0.05) given that the 2-haplotype (22) has frequency $T_{22} \geq t$ (by default $t = 0.001$). Similarly, we require that the expected number of false positive forbidden pairs be less than 0.05.

$$Pr(x \leq O_{22} | T_{22} \geq t) \leq \sum_{i=0}^{O_{22}} \binom{n}{i} t^i (1-t)^{n-i}$$

$$\leq \frac{0.05}{\binom{L}{2}} \tag{3.2}$$

Pairs of SNVs passing this linkage test (3.2) are classified as a *forbidden* SNV pairs.

Step 2: Constructing the SNV graph. The SNV graph $G = (V, E)$ consists of vertices corresponding to minor variants and edges corresponding to linked pairs of minor variants from different positions. If the intra-host population consists of very similar haplotypes, then graph G is very sparse. Indeed, the PacBio dataset for Influenza A virus encompassing $L = 2,500$ positions is split into 10,000 vertices, while the SNV graph contains only 700 edges, and, similarly, the simulated Illumina read dataset for the same haplotypes contains only 368 edges.

Note that the isolated minor variants correspond to genotyping errors unless they have a significant frequency. This fact allows us to estimate the number of errors per read, assuming that all isolated SNVs are errors. As expected, the distribution of the PacBio reads has a heavy tail (see Figure 3.2), which implies that most reads are (almost) error free, while a small number of heavy-tail reads accumulate most of the errors. Our analysis allows the identification of such reads, which can then be filtered out. By default, we filter out $\approx 10\%$ of PacBio reads, but we do not filter out any Illumina reads. The SNV graph is then constructed for the reduced set of reads. Such filtering allows the reduction of systematic errors and refines the SNV graph significantly.

Step 3: Finding cliques in the SNV graph G. Although the MAX CLIQUE is a well-known NP-complete problem and there may be an exponential number of maximal cliques in G, a standard Bron-Kerbosch algorithm requires little computational time since G is very sparse[25].

Step 4: Merging cliques in the clique graph C_G. The clique graph $C_G = (C, F, L)$ consists of

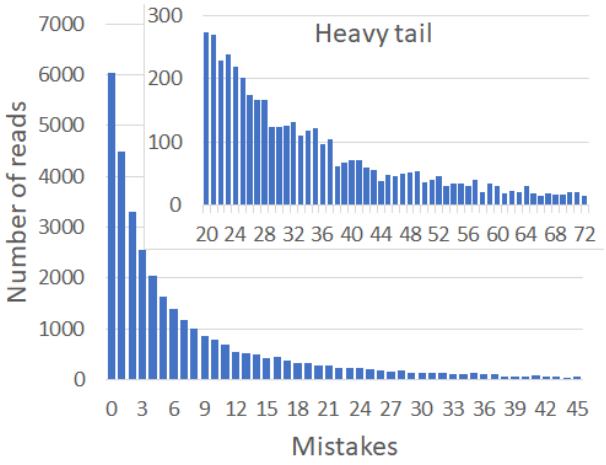

Figure 3.2 A typical distribution of errors in PacBio reads. The heavy tail indicates that a significant portion of errors is accumulated by a relatively small number of reads.

vertices corresponding to cliques in the SNV graph G and two sets of edges F and L. A *forbidding edge* $(p, q) \in F$ connects two cliques p and q with at least one forbidden pair of minor variants from p and q respectively. A *linking edge* $(p, q) \in L$ connects two cliques p and q, $(p, q) \notin F$, with at least one linked pair of minor variants from p and q respectively. Any true haplotype corresponds to a maximal L-connected subgraph H of C_G that does not contain any forbidding edge (see Fig. 3.1 (4)).

Unfortunately, even deciding whether there is a L-path between p and q avoiding forbidding edges is known to be NP-hard[98]. We find all subgraphs H as follows (see Fig. 3.3): (i) connect all pairs of vertices except connected with forbidding edges, (ii) find all maximal super-cliques in the resulted graph $C'_G = (C, C^{(2)} - F)$ using[25], (iii) split each super-clique into L-connected components, and (iv) filter out the L-connected components which are proper subsets of other maximal L-connected components.

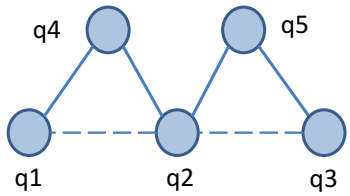

Figure 3.3 The clique graph C_G with 5 vertice corresponding to cliques in G, 4 edges and two forbidden pairs (q_1, q_2) and (q_2, q_3). There 3 maximal connected subgraphs avoiding forbidden pairs: $\{q_1, q_4\}$ $\{q_4, q_2, q_5\}$ $\{q_5, q_3\}$

Step 5: Partitioning reads between merged cliques and finding consensus haplotypes. Let S be the set of all positions containing at least one minor variant in V. Let q_S be an *major clique* corresponding to a haplotype with all major variants in S. The distance between a read r and a clique q equals the number of variants in q that are different from the corresponding nucleotides in r. Each read r is assigned to the closest clique q (which can possibly be q_S). In case of a tie, we assign r to all closest cliques.

Finally, for each clique q, CliqueSNV finds the consensus $v(q)$ of all reads assigned to q. Then $v(q)$ is extended from S to a full-length haplotype by setting all non-S positions to major SNVs.

Step 6: Estimating haplotype frequencies by expectation-maximization (EM) algorithm.

CliqueSNV estimates the frequencies of the assembled intra-host haplotypes via an expectation-maximization algorithm similar to the one used in IsoEM[124]. The algorithm starts by assigning equal frequencies to each haplotype and iteratively updates the probabilities to see observed data given the previous estimation of frequencies. Let K be the number of assembled viral variants, and let α be the probability of sequencing error. EM algorithm works as follows:

1. Initialize frequencies of viral variants $f_j^{(0)} \leftarrow \frac{1}{K}$,

Compute the probability of l_i-long read r_i $i = \overline{1, N}$, being emitted by viral variant $j = \overline{1, K}$,

$h_{ji} = \prod_{l=1}^{l_i}((1-\alpha)M_{ji,l} + \frac{\alpha}{3}(1 - M_{ji,l}))$,

where $M_{ji,l}$ - indicator if i-th read coincides with j-th viral variant in the position l

2. (Expectation) Update the amount of read r_i emitted by the jth viral variant $p_{ij} \leftarrow \frac{f_j^{(n-1)} h_{ji}}{\sum_{u=1}^{k} f_u^{(n-1)} h_{ui}}$

3. (Maximization) Update the frequency of the jth viral variant $f_j^{(n)} \leftarrow \frac{\sum_{i=1}^{N} p_{ij}}{\sum_{u=1}^{k}\sum_{i=1}^{N} p_{iu}}$

4. if $||f_j^{(n-1)} - f_j^{(n)}|| > \varepsilon$, then $n \leftarrow n + 1$ and go to step 2

5. Output estimated frequencies $f^{(n)}$

3.2 Results

3.2.1 Intra-host viral population sequencing benchmarks

We tested CliqueSNV's ability to assemble haplotype sequences and estimate their frequencies from PacBio and MiSeq reads using four real (experimental) and two simulated datasets from HIV and IAV samples (Table 3.1). Datasets contain two to ten haplotypes with frequencies 0.1 to 50%. The hamming distances between pairs of variants for each dataset are shown in Figure 3.4.

Name	Type	Virus	#haplotypes	Haplotype frequencies	Hamming distance
HIV9exp	experimental	HIV-1	9	0.2-50%	0.22-2.1%
HIV2exp	experimental	HIV-1	2	50-50%	1.2%
HIV5exp	experimental	HIV-1	5	20-20%	2-3.5%
IAV10exp	experimental	IAV	10	0.1-50%	0.1-1.1%
HIV7sim	simulated	HIV-1	7	14.3-14.3%	0.6-3%
IAV10sim	simulated	IAV	10	0.1-50%	0.1-1.1%

Table 3.1 Four experimental and two simulated sequencing datasets of human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). The datasets contain MiSeq and PacBio reads from intra-host viral populations consisting of two to ten variants each with frequencies in the range of 0.1-50%, and Hamming distances between variants in the range of 0.1-3.5%.

Experimental datasets:

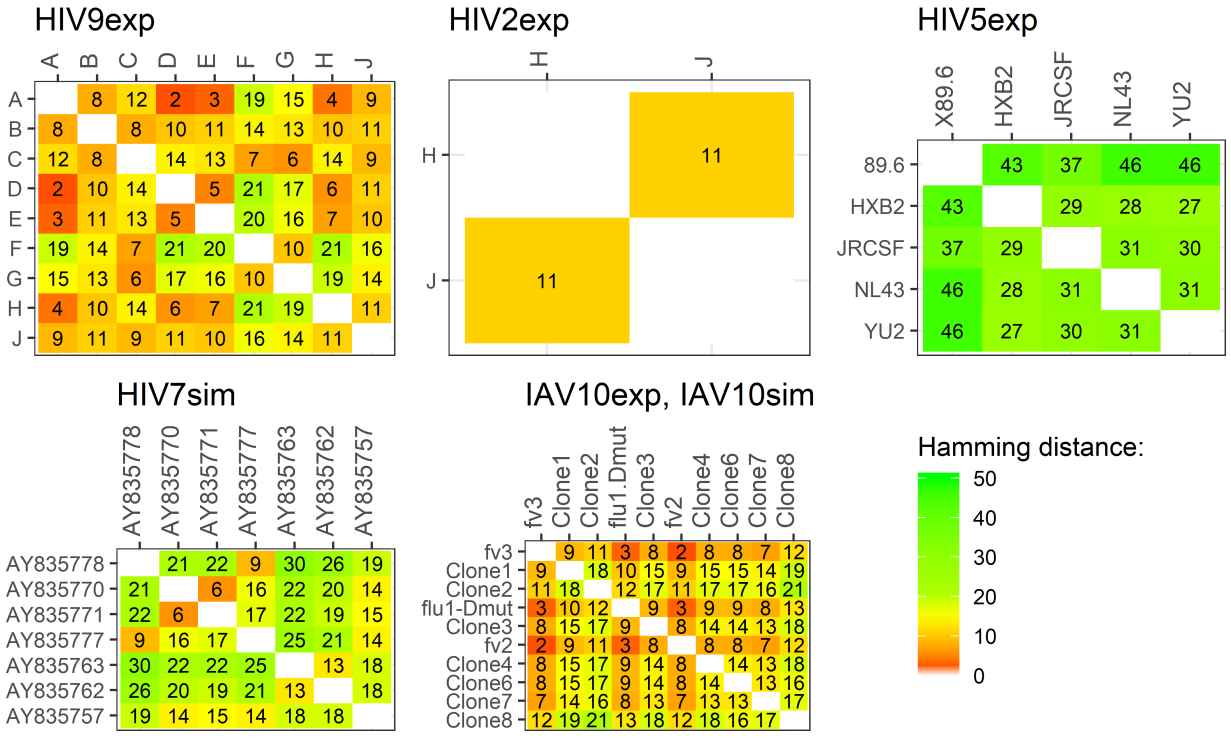

Figure 3.4 Pairwise hamming distances between variants in datasets HIV9exp, HIV2exp, HIV5exp, HIV7sim, IAV10sim, and IAV10exp.

1–2. *HIV-1 subtype B plasmid mixtures and MiSeq reads (HIV2exp and HIV9exp).* We designed nine *in silico* plasmid constructs comprising a 950-bp region of the HIV-1 polymerase *(pol)* gene that were then synthesized and cloned into pUCIDT-Amp (Integrated DNA Technologies, Skokie, IL). Each clone was confirmed by Sanger sequencing. This region at the beginning of *pol* can contain known protease and reverse transcriptase drug-resistant mutations and is monitored with sequence analysis for patient care. Each of these plasmids contains a specific set of point mutations chosen using mutation profiles from a real clinical study[190] to create nine unique synthetic HIV-1 *pol* haplotypes. Different proportions of these plasmids were mixed and then sequenced using an Illumina MiSeq protocol to obtain 2x300-bp reads

(see Supplementary Methods). HIV2exp is a mixture of two variants, and HIV9exp is a mixture of nine.

3. *HIV-1 subtype B mixture and MiSeq reads (HIV5exp).* This dataset consists of Illumina MiSeq 2×250-bp reads with an average read coverage of ~20,000× obtained from a mixture of five HIV-1 isolates: 89.6, HXB2, JRCSF, NL43, and YU2 available at[67]. Isolates have pairwise Hamming distances in the range from 2-3.5%(27 to 46-bp difference). The original HIV-1 sequence length was 9.3Kb, but was reduced to the beginning of *pol* with length of 1.3Kb.

4. *Influenza A mixture and PacBio reads (IAV10exp).* This benchmark contains ten influenza A virus clones that were mixed at a frequency of 0.1-50%. The Hamming distances between clones ranged from 0.1-1.1% (2-22–bp differences)[9]. The 2kb-amplicon was sequenced using the PacBio platform yielding a total of 33,558 reads of an average length of 1973 nucleotides.

Simulated datasets:

1. *HIV-1 subtype B mixture and MiSeq reads (HIV7sim).* This benchmark contains simulated Illumina MiSeq reads with 10k-coverage of 1-kb *pol* sequences. The reads were simulated from seven equally distributed HIV-1 variants chosen from the NCBI database: AY835778, AY835770, AY835771, AY835777, AY835763, AY835762, and AY835757. The Hamming distances between clones are in the range from 0.6-3.0%(6 to 30-bp differences). We used SimSeq[20] for generating reads.

2. *Influenza A mixture and MiSeq reads (IAV10sim).* This benchmark contains simulated IAV Illumina MiSeq reads with the same IAV haplotypes and their frequencies as for the IAV10exp benchmark. The sequencing of a 2kb-amplicon with 40k coverage with paired Illumina MiSeq reads was simulated by SimSeq[20] with the default sequencing error profile in SimSeq.

3.2.2 Validation metrics for viral population inference

3.2.2.1 Precision and recall

The quality of inference is usually measured by precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where TP is a number of true predicted haplotypes, FP is a number of false predicted haplotypes, and FN a number of undiscovered haplotypes.

Initially we measured precision and recall strictly by treating a predicted haplotype with a single mismatch as a FP. Additionally, like in[133] we introduced an acceptance threshold, which is a number of mismatches permitted for in a predicted haplotype to count as a TP.

3.2.2.2 Matching errors between populations

Unfortunately, precision and recall do not take into account (i) distances between true and inferred viral variants as well as (ii) the frequencies of the true and inferred viral variants. Instead, we

propose to use analogues of precision and recall defined for populations as follows.

Let $T = \{(t, f_t)\}$, be the true haplotype population, where f_t is the frequency of the true haplotype t, $\sum_{t \in T} f_t = 1$. Similarly, let $P = \{(p, f_p)\}$, be the reconstructed haplotype population, where f_p is the frequency of the reconstructed haplotype p, $\sum_{p \in P} f_p = 1$. Let d_{pt} be the edit distance between haplotypes p and t. Thus, instead of precision, we propose to use the *matching error* $E_{T \to P}$ measuring how well each reconstructed haplotype $p \in P$ weighted by its frequency is matched by the closest true haplotype.

$$E_{T \to P} = \sum_{p \in P} f_p \min_{t \in T} d_{pt}$$

Indeed, precision increases while $E_{T \to P}$ decreases and reaches 100% when $E_{T \to P} = 0$. Similarly, instead of recall, we propose to use the *matching error* $E_{T \leftarrow P}$ measuring how well each true haplotype $t \in T$ weighted by its frequency is matched by the closest reconstructed haplotype.[66]

$$E_{T \leftarrow P} = \sum_{t \in T} f_t \min_{p \in P} d_{pt}$$

Note that recall increases while $E_{T \leftarrow P}$ decreases and reaches 100% when $E_{T \leftarrow P} = 0$.

3.2.2.3 Earth mover's distance (EMD) between populations

The matching errors introduced above match haplotypes of true and reconstructed populations but do not match their frequencies. In order to simultaneously match haplotype sequences and their frequencies, we need to allow a fractional matching when portions of a single haplotype p of

population P are matched to portions of possibly several haplotypes of T and *vice versa*. Thus, we should separate f_p into f_{pt}'s each denoting portion of p matched to t such that $f_p = \sum_{t \in T} f_{pt}$, $f_{pt} \geq 0$. Symmetrically, f_t's are also separated into f_{pt}'s, i.e, $\sum_{p \in P} f_{pt} = f_t$. Finally, we should choose f_{pt}'s minimizing the total error of matching T to P also known as Wasserstein metric or EMD between T and P[111,102].

$$EMD(T, P) = \min_{f_{pt}>0} \sum_{t \in T} \sum_{p \in P} f_{pt} d_{pt}$$

$$\text{s.t.} \sum_{t \in T} f_{pt} = f_p, \text{ and} \sum_{p \in P} f_{pt} = f_t$$

EMD is efficiently computed as an instance of the transportation problem using network flows.

It is not surprising that EMD varies a lot over different benchmarks. Different benchmarks may have different complexity, which depends on the number of true variants, the frequency distribution, the similarity between haplotypes, sequencing depth, sequencing error rate, and many other parameters. We propose to measure the complexity of a benchmark as the EMD between the true population and a population consisting of a single consensus haplotype[185].

3.2.3 Performance of haplotyping methods

We compared CliqueSNV to 2SNV, PredictHaplo, and aBayesQR. Since CliqueSNV, PredictHaplo and aBayesQR can handle Illumina reads, we compared them on HIV9exp, HIV2exp, HIV5exp, HIV7sim, and IAV10sim datasets. Since CliqueSNV, 2SNV, and PredictHaplo can handle PacBio reads, we compared them on the IAV10exp dataset. We also used consensus sequences in the

comparison[185] because of its simplicity and to evaluate sequences most similar to those generated by the Sanger method[93].

The precision and recall of haplotype discovery for each method is provided in Table 3.2. CliqueSNV has the best precision and recall for five of the six datasets. For the HIV5exp dataset, PredictHaplo is more conservative and predicted less false positive variants (better precision) than CliqueSNV but the recall is the same for both methods. CliqueSNV has 100% precision and recall for three datasets, including the HIV2exp and IAV10exp experimental datasets and the HIV7sim simulated dataset.

Benchmark	CliqueSNV		aBayesQR		PredictHaplo	
	Precision	Recall	Precision	Recall	Precision	Recall
HIV9exp	**0.50**	**0.33**	0.08	0.11	0.00	0.00
HIV2exp	**1.00**	**1.00**	0.08	0.50	0.33	0.50
HIV5exp	0.50	**0.60**	0.00	0.00	**0.75**	**0.60**
HIV7sim	**1.00**	**1.00**	0.43	0.43	0.00	0.00
IAV10sim	**0.70**	**0.70**	0.13	0.10	0.33	0.10

(a)

Benchmark	CliqueSNV		2SNV		PredictHaplo	
	Precision	Recall	Precision	Recall	Precision	Recall
IAV10exp	**1.00**	**1.00**	0.82	0.90	0.70	0.70

(b)

Table 3.2 Prediction statistics of haplotype reconstruction methods using experimental and simulated (a) MiSeq and (b) PacBio data. The precision and recall was evaluated stringently such that if a predicted haplotype has at least one mismatch to its closest answer, then that haplotype is scored as a false positive.

Following Prabhakaran's study[133] we introduced an acceptance threshold, which is the number of mismatches permitted for a predicted haplotype to count as a TP. We report the numbers TP and FP for acceptance allowing from 0 to 30 mismatches (see Figure 3.5).

Matching distance analysis on Figure 3.6 shows that matching distances $E_{T\leftarrow P}$ and $E_{T\rightarrow P}$ are better for CliqueSNV than for both PredictHaplo and aBayesQR on all MiSeq datasets. Using

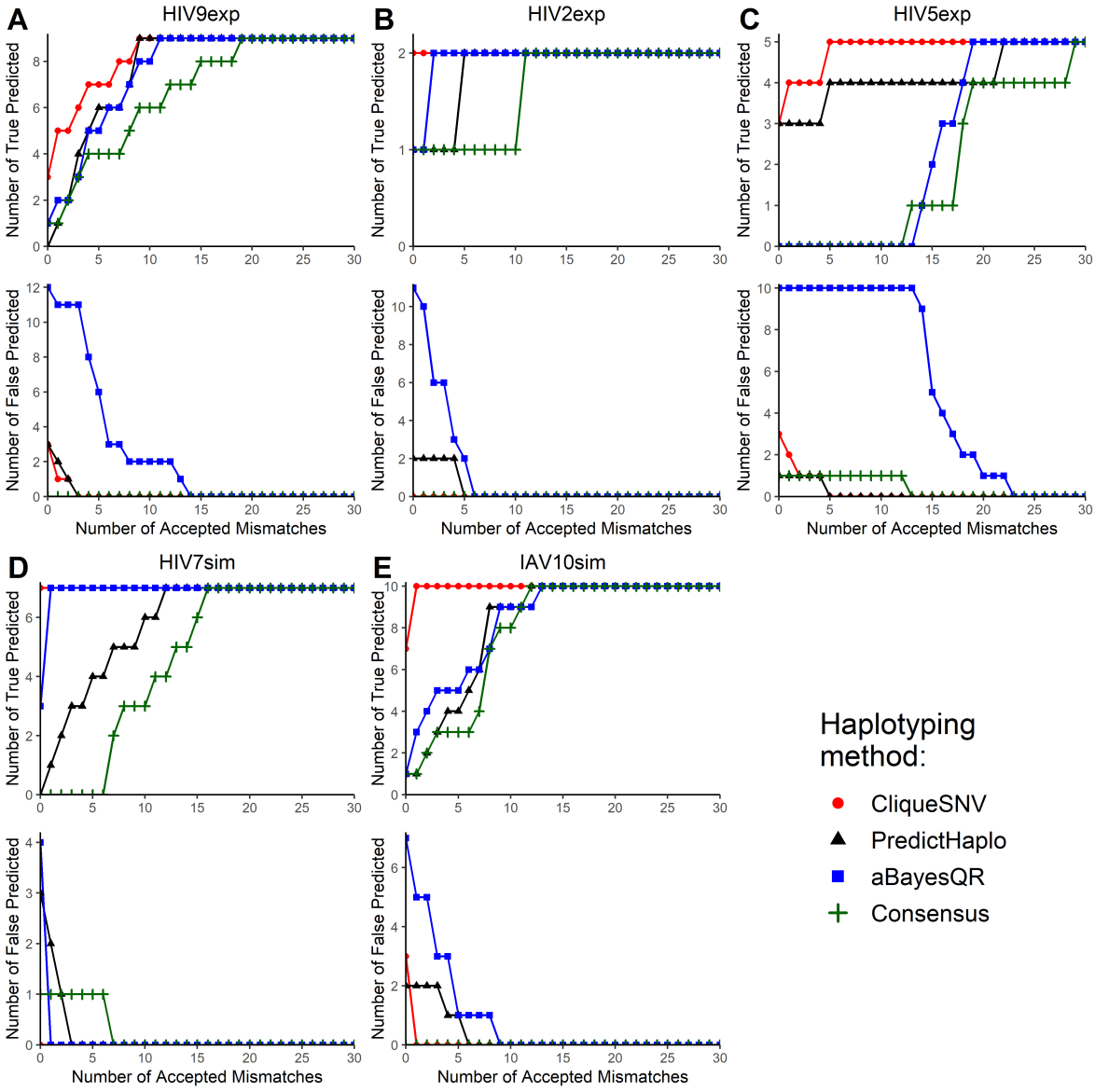

Figure 3.5 The number of true and false predicted haplotypes depending on the number of accepted mismatches for five benchmarks: (A) HIV9exp; (B) HIV2exp; (C) HIV5exp; (D) HIV7sim; (E) IAV10sim. Two haplotypes are regarded identical if the Hamming distance between them is at most the number of accepted mismatches.

HIV2exp, HIV7sim, and IAV10sim datasets, the $E_{T\leftarrow P}$ and $E_{T\rightarrow P}$ for CliqueSNV are very close to zero indicating that the predictions are almost perfect. Since $E_{T\leftarrow P}$ and $E_{T\rightarrow P}$ correlate with precision and recall, matching distance analysis indicates that CliqueSNV has a better precision, and

significantly outperformed both PredictHaplo and aBayesQR. Since aBayesQR has higher $E_{T\leftarrow P}$ on MiSeq datasets, it is more likely to make more false predictions. Notably, on the HIV7sim dataset, aBayesQR outperformed PredictHaplo and was just a little behind CliqueSNV.

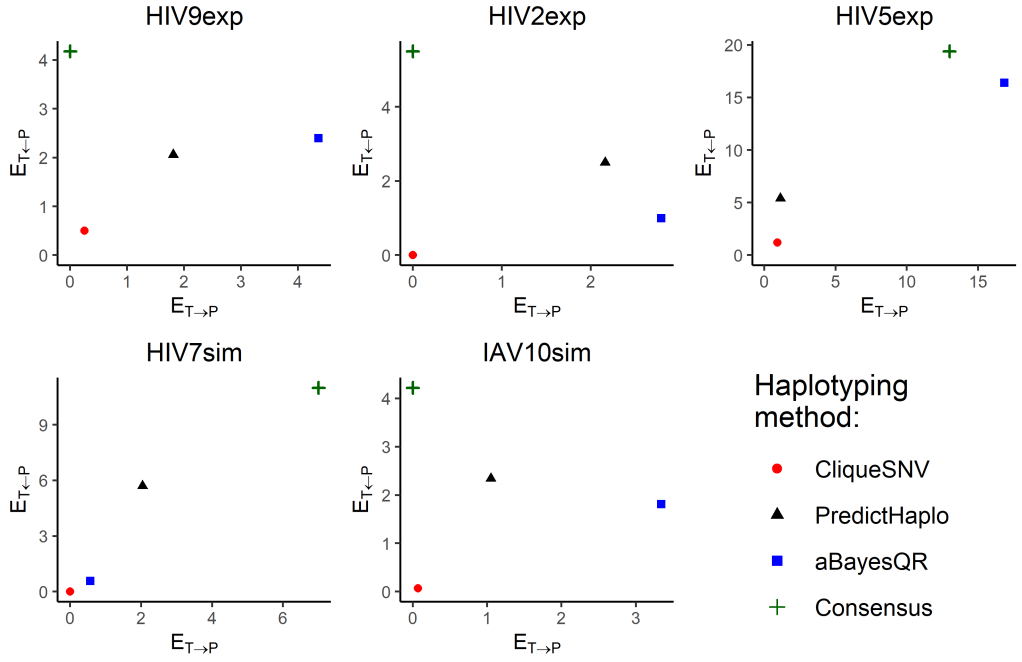

Figure 3.6 Matching distances $E_{T\leftarrow P}$ and $E_{T\rightarrow P}$ between a true haplotype population and a reconstructed haplotype population for five benchmark datasets for human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). Matching distance $E_{T\leftarrow P}$ is shown on the x-axis and $E_{T\rightarrow P}$ is shown on the y-axis for each benchmark. Smaller matching distances indicate better approximation of a true haplotype population T by a reconstructed haplotype population P. Haplotype populations were reconstructed with CliqueSNV, aBayesQR, PredictHaplo and a population consisting of a single consensus haplotype.

Figure 3.7 shows the EMD distance between predicted and true haplotype populations for all five MiSeq datasets. The exact EMD values are provided in Table 3.3. CliqueSNV has a lower (better) EMD than other tools on all benchmarks. Using the simulated and PacBio datasets, CliqueSNV has almost zero EMD indicating almost ideal predictions. PredictHaplo has a lower EMD than

aBayesQR on four out of five MiSeq datasets. aBayesQR has almost zero-EMD on HIV7sim and is only slightly behind CliqueSNV, while on HIV5exp, aBayesQR performs significantly worse than the other methods.

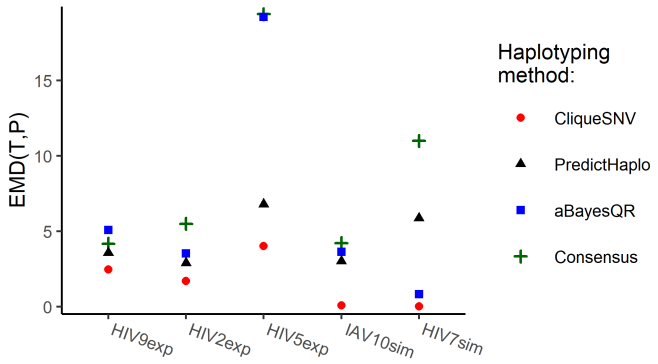

Figure 3.7 Earth Movers' Distance (EMD) between true and reconstructed haplotype populations. Four haplotyping methods (CliqueSNV, aBayesQR, PredictHaplo, Consensus) are benchmarked using three experimental and two simulated datasets for human immunodeficiency virus type 1 (HIV-1) and influenza A virus (IAV). For all benchmarks the CliqueSNV predictions are the closest to the true populations.

Tables 3.4 and 3.5 describe the true variant IDs and their frequencies datasets, respectively, and report for each true variant T the quality of its prediction: the edit distance to the closest predicted variant (Err), and the frequency of the closest predicted variant (PF). The row EMD reports the EMD distance from the population of the true variants to the consensus (underscored) and to the population of variants predicted by the corresponding method. Note that the EMD to the consensus is a measure of the benchmark diversity.

CliqueSNV, 2SNV, and PredictHaplo were compared on the IAV10exp benchmark dataset (see Table 3.5. CliqueSNV correctly recovered all 10 true variants, including the haplotype with frequencies significantly below the error rate. 2SNV recovered nine true variants but reports one false

Benchmark	Consensus	CliqueSNV		aBayesQR		PredictHaplo	
	EMD	EMD	Improvement	EMD	Improvement	EMD	Improvement
HIV9exp	4.18	**2.47**	**40.83 %**	5.09	-21.85 %	3.58	14.30 %
HIV2exp	5.50	**1.71**	**68.95 %**	3.53	35.80 %	2.91	47.08 %
HIV5exp	19.40	**4.03**	**79.20 %**	19.22	0.91 %	6.80	64.97 %
HIV7sim	11.00	**0.02**	**99.84 %**	0.84	92.34 %	5.87	46.68 %
IAV10sim	4.22	**0.09**	**97.77 %**	3.64	13.73 %	3.03	28.15 %
Mean Improvement			**77.32 %**		24.19 %		40.23 %

(a)

Benchmark	Consensus	CliqueSNV		2SNV		PredictHaplo	
	EMD	EMD	Improvement	EMD	Improvement	EMD	Improvement
IAV10exp	4.22	**0.22**	**94.69%**	0.23	94.46%	0.38	91.02%

(b)

Table 3.3 Earth Movers' Distance from predicted haplotypes to the true haplotype population and haplotyping method improvement. Four haplotyping methods(aBayesQR, CliequeSNV, Consensus, PredictHaplo) are benchmarked on five MiSeq datasets (a) and IAV10exp dataset (b). The improvement shows how much better is prediction of haplotyping method over inferred consensus, and it is calculated as $\frac{(EMD_c - EMD_m)\times 100\%}{EMD_c}$, where EMD_c is an EMD for consensus, and EMD_m is an EMD for method. CliqueSNV outperformed all other methods in accuracy on all datasets.

positive. PredictHaplo recovered only seven true variants and falsely predicted three variants. To further explore the precision of these three methods with the IAV10exp data, we simulated low-coverage datasets by randomly subsampling $n = 16K, 8K, 4K$ reads from the original data (see Table 3.6). For each dataset, CliqueSNV found at least one true variant more than both 2SNV and PredictHaplo.

Finally, Table 3.7 reports the performance of three methods on full-length genomes. We normalize EMD over the genomic length so that the resulted EMD are in the same range and can be compared for different genomic regions. On average, CliqueSNV for all lower bounds on frequency (2%, 5%, and 10%) outperforms PredictHaplo, but for 2 out of 4 full-length benchmarks PredictHaplo is more accurate than CliqueSNV.

HIV9exp		CliqueSNV			PredictHaplo			aBayesQR		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
A	50	1	32.2	0	1	45.2	1	1	12.9	4
B	25	2	14.5	1	2	25.9	3	2	13	0
C	13	3	28.9	0	3	28.9	2	3	4.13	3
D	6.3	4	19.1	0	1		3	4	14.5	1
E	3.2	1		3	1		4	1		4
F	1.6	3		7	3		9	3		3
G	0.8	5	2.98	1	3		8	3		3
H	0.4	1		4	1		5	1		4
J	0.2	1		9	3		9	5	12.1	5
EMD	4.18			2.47			3.58			5.09

HIV2exp		CliqueSNV			PredictHaplo			aBayesQR		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
H	50	1	34.5	0	1	18.3	5	1	9.75	2
J	50	2	65.5	0	2	56.8	0	2	10.75	0
EMD	5.5			1.71			2.91			3.53

HIV5exp		CliqueSNV			PredictHaplo			aBayesQR		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
89.6	20	1	12.5	0	1	21.8	0	1	9.94	18
HXB2	20	2	6.9	5	2		22	2	9.08	15
JRCSF	20	3	7.55	1	3	29	0	3	8.16	14
NL43	20	4	16.9	0	4	26.6	0	4	7.36	16
YU2	20	5	10.8	0	2	22.7	5	4		19
EMD	19.4			6.52			6.8			19.2

HIV7sim		CliqueSNV			PredictHaplo			aBayesQR		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
AY835778	14.3	1	14.3	0	1	39	7	1	14.4	1
AY835770	14.3	2	14.3	0	2		5	2	15.1	1
AY835771	14.3	3	14.3	0	2	28.7	1	3	12.1	1
AY835777	14.3	4	14.3	0	1		2	4	15.5	1
AY835763	14.3	5	14.3	0	3	32.3	3	5	14.3	0
AY835762	14.3	6	14.2	0	3		10	6	14.4	0
AY835757	14.3	7	14.3	0	1		12	7	14.2	0
EMD	11			0.018			5.87			0.84

IAV10sim		CliqueSNV			PredictHaplo			aBayesQR		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
fv3	50	1	50.1	0	1	76.3	0	1	35.2	1
Clone1	25	2	24.9	0	2	18.5	4	2	14	1
Clone2	13	3	12.4	0	3	5.27	6	3	8.11	6
flu1-Dmut	6.3	4	6.3	1	1		3	1		2
Clone3	3	5	3.1	0	1		8	4	4.24	0
fv2	1.6	6	1.6	0	1		2	1		3
Clone4	0.8	7	0.78	1	1		8	1		9
Clone6	0.4	8	0.41	0	1		8	1		9
Clone7	0.2	9	0.2	1	1		7	1		8
Clone8	0.1	10	0.1	0	1		12	1		13
EMD	4.22			0.0939			3.03			3.64

TV - id of a true variant, TF - frequency of the true variant in a mixture, PV - id of the closest predicted variant to the true variant, PF - frequency of the closest predicted variant, Err - number of mismatches between the true variant and the predicted variant. The underscored value is the EMD distance to the population consisting of a single variant coinciding with the read consensus.

Table 3.4 Comparison of CliqueSNV with PredictHaplo and aBayesQR on simulated and real Illumina data

IAV10exp		CliqueSNV			2SNV			PredictHaplo		
TV	**TF, %**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**	**PV**	**PF, %**	**Err**
fv3	50	1	52.6	0	1	51.8	0	1	56.7	0
Clone1	25	2	23.7	0	2	23.7	0	2	23.7	0
Clone2	13	3	12.6	0	3	12.5	0	3	13.7	0
flu1-Dmut	6.3	4	6.41	0	4	6.39	0	4	6.01	0
Clone3	3	5	2.32	0	5	2.3	0	5	3.01	0
fv2	1.6	6	1.17	0	6	1.19	0	1		2
Clone4	0.8	7	0.69	0	7	0.7	0	6	2.9	0
Clone6	0.4	8	0.35	0	8	0.34	0	7	1.2	0
Clone7	0.2	9	0.12	0	9	0.12	0	1		7
Clone8	0.1	10	0.05	0	1		12	1		12
EMD	4.22			0.22			0.23			0.38

TV - id of a true variant, TF - frequency of the true variant in a mixture, PV - id of the closest predicted variant to the true variant, PF - frequency of the closest predicted variant, Err - number of mismatches between the true variant and the predicted variant. The underscored value is the EMD distance to the population consisting of a single variant coinciding with the read consensus.

Table 3.5 Comparison of CliqueSNV with PredictHaplo and 2SNV on IAV10exp

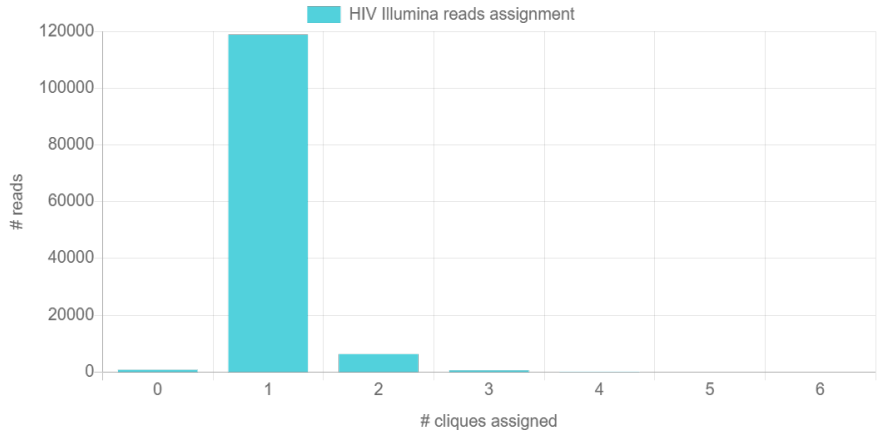

Figure 3.8 The number of reads assigned to different number of cliques in HIV Illumina dataset.

3.2.4 Runtime comparison

For comparing of running time of each method, we used the same PC (Intel(R) Xeon(R) CPU X5550 2.67GHz x2 8 cores per CPU, DIMM DDR3 1,333 MHz RAM 4Gb x12) with the CentOS 6.4 operating system. The runtime of CliqueSNV is sublinear with respect to the number of reads while the runtime of PredictHaplo and 2SNV exhibit super-linear growth. For the 33k IAV10sim reads the CliqueSNV analysis took 21 seconds, while PredictHaplo and 2SNV took around 30

# of PacBio Reads	Method	Variant	fv3	Clone1	Clone2	flu1-Dmut	Clone3	fv2	Clone4	Clone5	Clone6	Clone7	FP
		True Freq.,%	50	25	12.5	6.25	3.125	1.56	0.78	0.39	0.19	0.097	
33.5K (all)	CliqueSNV	Match	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	0
		Freq., %	52.6	23.7	12.6	6.4	2.3	1.17	0.7	0.35	0.12	0.051	0
	2SNV	Match	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	1
		Freq., %	51.8	23.7	12.5	6.4	2.3	1.2	0.7	0.3	0.1	0	1.0
	PredictHaplo	Match	✓	✓	✓	×	✓	×	✓	✓	×	×	0
		Freq.,%	56.7	23.8	13.7	0	3.1	0	1.5	1.2	0	0	0
Subsampling													
16K	CliqueSNV	Match,%	100	100	100	100	100	90	100	100	100	20	0.1
		Freq., %	52.9	23.7	12.5	6.4	2.3	1.19	0.71	0.32	0.12	0.69	1.15
	2SNV	Match,%	100	100	100	100	100	100	100	100	0	0	0.2
		Freq., %	52.4	23.7	12.5	6.4	2.3	1.1	0.7	0.3	0	0	0.6
	PredictHaplo	Match	100	100	100	70	100	0	100	40	0	0	0.3
		Freq.,%	54.2	23.5	13.1	6.0	2.9	0	1.4	1.0	0	0	0.5
8K	CliqueSNV	Match,%	100	100	100	100	100	90	100	100	30	0	0
		Freq., %	52.8	23.6	12.5	6.5	2.3	1.2	0.7	0.35	0.16	0	0
	2SNV	Match,%	100	100	100	100	100	100	100	0	0	0	0
		Freq., %	53.1	23.7	12.5	6.5	2.3	1.25	0.7	0	0	0	0
	PredictHaplo	Match,%	100	100	100	0	100	0	100	20	0	0	0.2
		Freq.,%	58.1	24.0	12.7	0	3.1	0	1.6	1.3	0	0	0.5
4K	CliqueSNV	Match,%	100	100	100	100	100	80	100	40	0	0	0
		Freq., %	53.3	23.7	12.3	6.4	2.4	1.19	0.7	0.39	0	0	0
	2SNV	Match,%	100	100	100	100	100	100	20	0	0	0	0
		Freq., %	53.7	23.7	12.3	6.5	2.4	1.2	0.9	0	0	0	0
	PredictHaplo	Match,%	100	100	100	0	70	0	10	0	0	0	0.3
		Freq.,%	60.1	23.9	12.8	0	3.5	0	2.5	0	0	0	0.5

Table 3.6 Comparison of CliqueSNV, 2SNV and PredictHaplo on full and sub-sampled data ***(PacBio, experimental)***. For all 33.5K reads, the sign "✓" (respectively, "×") denotes fully matched (respectively, unmatched) true variant and the column FP reports the number of incorrectly predicted variants (false positives) and their total frequency. For each sub-sample size (16K,...,4K), the table reports the percent of runs when a variant is completely matched and its average frequency over runs when the variant was detected. Similarly, the column FP reports the average number of false positive variants and their average total frequency. Colors indicate the percent of matched variants: green - high percent, red - low percent.

minutes. The runtime of CliqueSNV is quadratic with respect to the number of SNVs rather than by the length of the sequencing region. For our next runtime comparison, we generated five HIV-1 variants within 1% Hamming distance from each other, which is the estimated distance between related HIV variants from the same person[178]. Then we simulated 1M Illumina reads for sequence regions of length 566, 1132, 2263 and 9181 nucleotides for which CliqueSNV required 37, 144, 227, and 614 seconds, respectively, for analyzing these datasets. CliqueSNV is significantly faster than aBayesQR and PredictHaplo. For example, using the HIV2exp benchmark the runtimes of

Benchmark	Length	Consensus	CliqueSNV			PredictHaplo	aBayesQR
			2%	5%	10%		
HCV10sim	1K	13.52	64.12	72.59	65.86	314.87	did not finish
	2K	13.85	169.16	133.06	108.46	972.41	did not finish
	5K	16.79	3666.76	3117.49	221.70	6472.83	did not finish
	full-length	15.27	3703.01	3559.10	483.77	58509.17	did not finish
ZIKV3sim	1K	34.61	81.88	91.21	91.50	88.76	4409.53
	2K	31.57	104.71	115.81	106.82	342.31	did not finish
	5K	33.62	161.90	156.31	160.64	1775.49	did not finish
	full-length	35.20	271.55	281.47	284.54	12114.49	did not finish
ZIKV15sim	1K	13.33	114.42	117.75	139.08	314.87	did not finish
	2K	13.16	148.40	153.95	147.76	342.31	did not finish
	5K	13.70	337.82	229.16	166.85	1775.49	did not finish
	full-length	13.66	10305.01	604.60	286.19	12114.49	did not finish
HIV5full	1K	21.60	247.84	215.70	208.73	155.11	24462.81
	2K	20.18	1282.03	460.03	374.76	459.40	28820.99
	5K	19.77	5291.37	1787.24	337.52	2982.96	did not finish
	full-length	20.26	8084.50	4970.50	1153.09	14404.43	did not finish
Average over all benchmarks		20.63	2127.16	1004.12	271.08	7071.21	21628.58

Table 3.7 Running time of performed experiments (seconds) for full-length benchmarks.

aBayesQR was over ten hours, PhedictHaplo took 24 minutes, while CliqueSNV only required 79 seconds (see Figures 3.9 and 3.10).

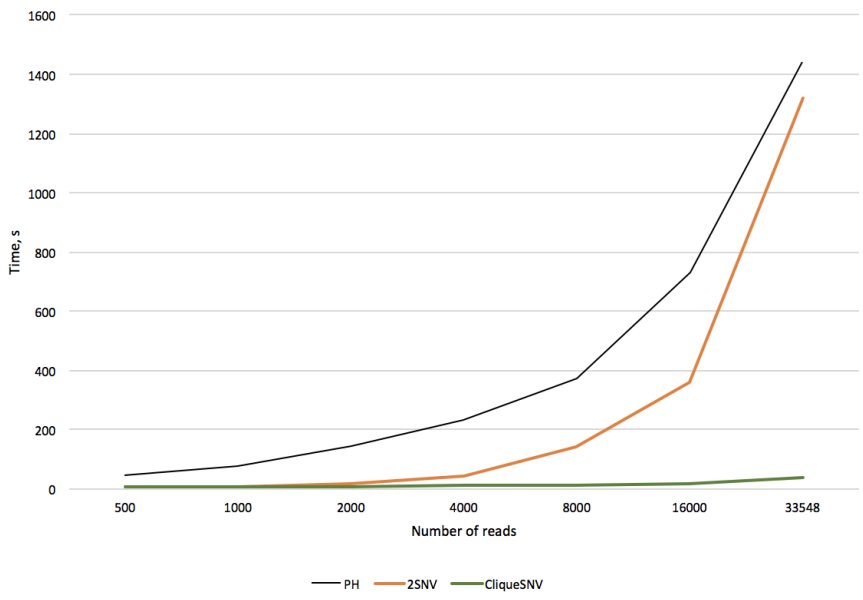

Figure 3.9 Runtime of PredictHaplo (PH), 2SNV and CliqueSNV on datasets with different sizes.

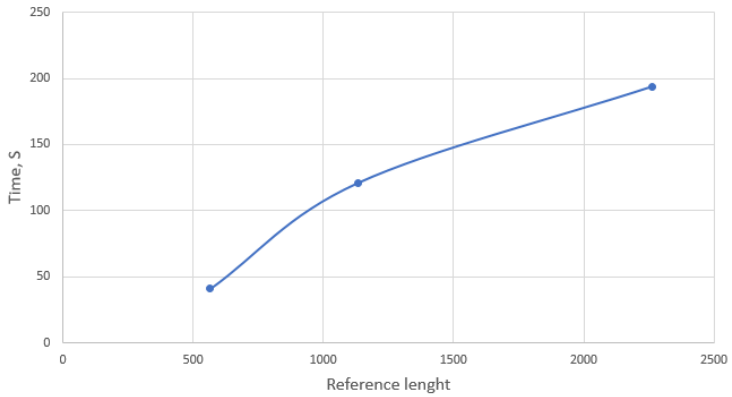

Figure 3.10 CliqueSNV runtime on datasets with different reference length and same coverage (about 1M reads in total).

3.3 Discussion

Assembly of haplotype populations from noisy NGS data is one of the most challenging problems of computational genomics. High-throughput sequencing technologies, such as Illumina MiSeq and HiSeq, provide deep sequence coverage that allows discovery of rare, clinically relevant haplotypes. However, the short reads generated by the Illumina technology require assembly that is complicated by sequencing errors, an unknown number of haplotypes in the samples, and the genetic similarity of haplotypes within a sample. Furthermore, the frequency of sequencing errors in Illumina reads is comparable to the frequencies of true minor mutations[159]. The recent development of single-molecule sequencing platforms such as PacBio produce reads that are sufficiently long to span entire genes or small viral genomes. Nonetheless, the error rate of single-molecule sequencing is exceptionally high and could reach $13 - 14\%$[135], which hampers PacBio sequencing to detect and assemble rare viral variants.

We developed CliqueSNV, a new reference-based assembly method for reconstruction of rare

genetically-related viral variants. CliqueSNV allows for accurate haplotyping in the presence of high sequencing error rates, which is also suitable for both single-molecule and short-read sequencing. In contrast to other haplotyping methods, CliqueSNV infers viral haplotypes by detection of clusters of statistically linked SNVs rather than through assembly of overlapping reads used with methods such as Savage[13]. Using experimental data, we demonstrate that CliqueSNV can detect haplotypes with frequencies as low as 0.1%, which is comparable to the precision of many deep sequencing-based point mutation detection methods[61,75]. Furthermore, CliqueSNV can successfully infer and reconstruct viral variants, which differ by only a few mutations, thus demonstrating the high precision of identifying closely related variants. Another significant advantage of CliqueSNV is its low computation time, which is achieved by a very fast searching of linked SNV pairs and the application of the special graph-theoretical approach to SNV clustering.

Although very accurate and fast, CliqueSNV has some limitations. Unlike Savage[13], CliqueSNV is not a *de novo* assembly tool and requires a reference viral genome. This obstacle could easily be addressed by using Vicuna[185] or other analogous tools to assemble a consensus sequence, which can then be used as a reference. Another limitation is for variants that differ only by isolated SNVs separated by long conserved genomic regions longer than the read length which may not be accurately inferred by CliqueSNV. While such situations usually do not occur for viruses, where mutations are typically densely concentrated in different genomic regions, we plan to address this limitation in the next version of CliqueSNV.

The ability to accurately infer the structure of intra-host viral populations makes CliqueSNV applicable for studying viral evolution, transmission and examining the genomic compositions of

RNA viruses. However, we envision that the application of our method could be extended to other highly heterogeneous genomic populations, such as metagenomes, immune repertoires, and cancer cell genes.

CHAPTER 4
INTER-HOST VIRAL ANALYSIS USING NGS

4.1 Introduction

Inferring transmission clusters, transmission directions, and sources of outbreaks from viral sequencing data are crucial for viral outbreaks investigation. Outbreaks of RNA viruses, such as Human Immunodeficiency Virus (HIV) and Hepatitis C virus (HCV), are particularly dangerous and pose a significant problem for public health. It is well known that genomes of RNA viruses mutate at extremely high rates[51]. As a result, RNA viruses exist in infected hosts as populations of closely related variants called quasispecies[45,47]. However, only recently with the progress of sequencing technologies, it became possible to identify and sample quasispecies at great depth[55,7,78,174,155,29]. Consequently, a contribution of sequencing technologies to molecular surveillance of viral disease epidemic spread becomes more and more substantial[178,179].

Computational methods can be used to infer transmission characteristics from sequencing data. The first question usually is whether two viral populations belong to the same outbreak. The methods typically utilize the simple observation that all samples from the same outbreak are genetically related, so they use some measure of genetic relatedness as a predictor for epidemiological relatedness[178,179,30]. The second question is which samples constitute isolated outbreaks. For this purposes, we define a transmission cluster as a connected set of genetically related viral populations. The third questions we address in this chapter is ”Who is the source of infection?”. This questions is the most difficult to answer, and there were only a few attempts to do it computationally using solely genomic data[147] without invoking additional epidemiological information[41]. To

the best of our knowledge, there is still no freely available computational tool for this problem.

Computational methods for detection of viral transmissions and inference of transmission clusters are often consensus-based, i.e. they analyze only a single representative sequence per intra-host population (for example, consensus sequence). Such methods assign two hosts into one transmission cluster, if the distances between corresponding sequences do not exceed a predefined threshold[178,179]. Although consensus-based methods proved to be useful, they do not take into account intra-host viral diversity. Inclusion of whole intra-host populations into analysis is important, because minor viral variants are frequently responsible for transmission of RNA viruses[59,6].

Recently published computational approach (further referred to as MinDist)[30] uses the minimal genetic distance between sequences of two viral populations as a measure of genetic relatedness of intra-host viral populations. Since minimal genetic distances between different pairs of populations can be achieved on various pairs of sequences, this approach takes into account intra-host diversity.

However, both consensus-based and MinDist approaches have further limitations. First of all, they do not allow to detect directions of transmissions, which is crucial for detection of outbreak sources and transmission histories. Secondly, distance thresholds utilized by both approaches could be derived from analysis of limited or incomplete experimental data and highly data- and situation-specific, with different viruses or even different genomic regions of the same virus requiring specifically established thresholds.

In this chapter, we address the above limitations by proposing a novel algorithms $VOICE$ and an improvement of the MinDist algorithm. The new algorithms allow to infer important epidemiological characteristics, including genetic relatedness, directions of transmissions and transmission

clusters.

- Viral Outbreak InferenCE ($VOICE$) is a simulation-based method which imitates viral evolution as a Markov process in the space of observed viral haplotypes
- MinDistB method is a modification of MinDist, which takes into account the sizes of relative borders of each pair of viral populations.

The proposed methods were validated on the experimental data obtained from HCV outbreaks. Comparative results suggest that our methods are efficient in epidemiological characteristics inference.

4.2 Methods

4.2.1 Viral outbreak inference ($VOICE$) simulation method

$VOICE$ is an approach to predict epidemiological characteristics. It simulates the process of evolution from one viral population (source) into another (recipient) as a Markov process on a union of both populations. VOICE starts evolution from a subset of source sequences called the *border set* and estimates the number of generations required to acquire a genetic heterogeneity observed in the recepient.

Formally, given two sets of viral sequences P_1 and P_2, $VOICE$ simulates viral evolution to estimate times t_{12} and t_{21} needed to cover all sequences from the recipient population under the assumptions that first and second host were sources of infection. Based on the value $\min\{t_{12}, t_{21}\}$, the algorithm decides whether the populations are related. The direction of possible transmission between the related pair is assumed to follow the direction which requires less time.

The simulation starts from the *δ-border set* B_1, which contains viral variants that are likely the closest to variants transmitted between P_1 and P_2. It is defined as the set of vertices of P_1 minimizing pairwise Hamming distance D between vertices from P_1 and P_2 up to a constant δ:

$$B_1 = \{u \in P_1 : \exists v \in P_2 \;\; D(u,v) = \min_{x \in P_1, y \in P_2} D(x,y) + \delta\}$$

(see Fig. 4.3). The constant δ is a parameter, with the default value 1.

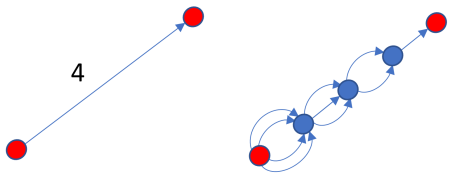

Figure 4.1 Edge subdividing

The simulated evolutionary process is carried out in the evolutionary space represented by the *variant graph* $G(B_1, P_2)$, which is constructed as follows. First, construct a union of all minimal spanning trees of the complete graph on a vertex set $B_1 \cup P_2$ with the edge weights equal to Hamming distances between variants (sometimes referred to as a pathfinder network $PFNet(n-1, \infty)$[138,28]). Then substitute every edge in graph with two directed edges of the same weight. Next, subdivide each edge (u_1, u_2) of weight $w \geq 2$ with $w-1$ vertices $v_1, ..., v_{w-1}$ and add multiple directed edges as follows: add $w-1$ edges between vertices u_1 and v_1; $w-2$ edges between v_1 and v_2; and so forth as shown on Figure 4.1. This model can be explained as follows: to mutate from vertex u_1 to u_2 during simulation, there should occur mutations at w positions that are different between u_1 and u_2. During the first step, simulation can mutate any of w positions, then any of

$w-1$ positions on the second step and so forth.

The simulation starts from all border vertices B_1 and runs until all the vertices of the population P_2 are reached. At the beginning of the simulation, border vertices get count equal to 1, and the rest of the vertices get count 0. Each tact simulates variants replication by updating vertex counts according to one of the three following scenarios happening with the specified probabilities (see Figure 4.2). First, if during replication there are no mutations, then the vertex v replicates itself and its count label is incremented. This happens with the probability p_1 (4.1). Second, the vertex can mutate into one of its neighboring vertices with probability p_2 (see (4.2)), in which case the count of the neighbor is incremented. Finally, with probability p_3, vertex does not produce any viable offspring, in which case vertex counts are not changed. If the count of a vertex reaches the maximum allowed variant population size C_{max}, then it is not increased. The probabilities of these scenarios are calculated as follows:

$$p_1 = (1-3\epsilon)^L \tag{4.1}$$

$$p_2 = p_1 \frac{\epsilon}{1-3\epsilon} \tag{4.2}$$

$$p_3 = 1 - p_1 - p_2 \deg^-(v) \tag{4.3}$$

where ϵ is the mutation rate, L is the genome length and $\deg^-(v)$ is an outdegree of a vertex v.

Algorithm 1 represents the flow of the method. The time t_{12} is computed as the average over s simulations. The same procedure is repeated for the opposite direction of the transmission with its border set B_2 and the time t_{21} is computed. The value $\min\{t_{12}, t_{21}\}$ determines which direction of

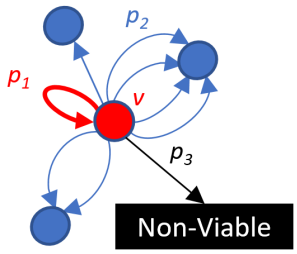

Figure 4.2 All possible moves of a vertex v

transmission is more likely.

Algorithm 1 $VOICE$ (Viral Outbreak InferenCE)

Require: Two sets of viral variants P_1, P_2.
Ensure: Time $t_{1,2}$ to evolve from P_1 to P_2.

find the δ-border set B_1
build the variant graph $G = G(B_1, P_2)$
$t \leftarrow 0$
Assign the number of copies $c_v^t \leftarrow 1$ to each variant $v \in B_1$ and $c_v^t \leftarrow 0$ to each variant $v \in P_2 \setminus B_1$
while there are variants $v \in P_2$ with $c_v^t = 0$ **do**
    $c_v^{t+1} \leftarrow c_v^t$ for every $v \in V(G)$
    for each variant $v \in V(G)$ **do**
        for $i = 1, ..., c_v^t$ **do**
            with a probability p_1, $c_v^{t+1} \leftarrow \min\{c_v^{t+1} + 1, C_{max}\}$
            with a probability p_2, $c_u^{t+1} \leftarrow \min\{c_u^{t+1} + 1, C_{max}\}$, where u is a randomly chosen neighbor of v
    $t \leftarrow t + 1$
$t_{1,2} \leftarrow t$

4.2.1.1 Data normalization

The sizes of observed intra-host viral populations may significantly vary due to sampling and sequencing biases. Since the larger population will require more time to cover, the estimation of t_{12} and t_{21} could be biased. VOICE avoids such biases by normalizing the intra-host population sizes. The deterministic normalization partitions each viral population into q clusters using hierar-

chical clustering and each cluster is replaced with the consensus of its members. The subsampling normalization randomly chooses q sequences from each population. The procedure is repeated r times, and the final result is an average over all subsamplings.

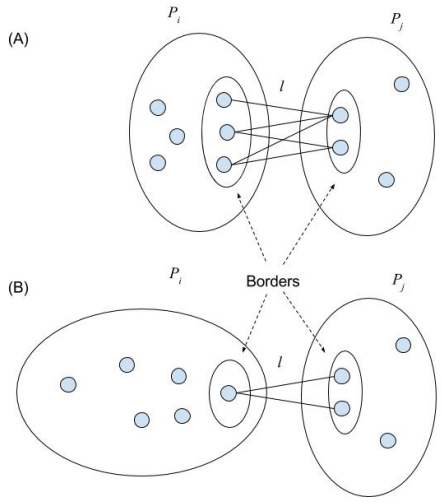

Figure 4.3 δ-Crossing between two viral populations P_1 and P_2 $l \leq d(u,v) + \delta$; (A) $|B_\delta| = 5$; (B) $|B_\delta| = 2$

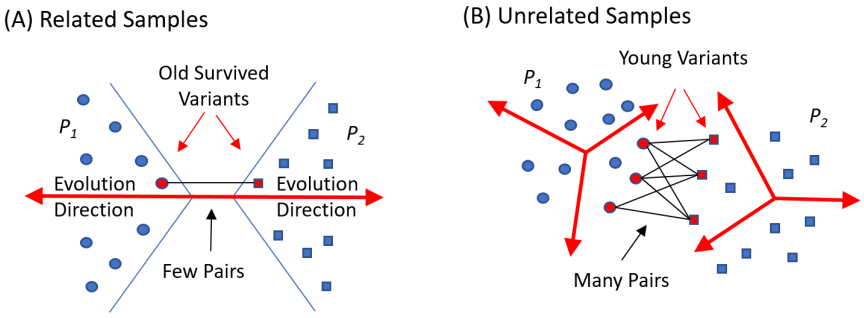

Figure 4.4 Intuition behind the MinDistB method. (A) Related samples – crossing is between old survived variants (B) Unrelated samples –crossing is between many young variants which are close to each other by chance.

4.2.1.2 Identification of genetic relatedness, transmission directions, clusters and sources of outbreaks

$VOICE$ produces a weighted directed genetic relatedness graph $G = (V, A, w)$ with $V = \mathcal{P}$. An arc P_iP_j is in A whenever populations P_i and P_j are genetically related, i.e., value $\min\{t_{ij}, t_{ji}\}$ is less than a threshold. Weakly connected components of G represent transmission clusters or outbreaks. To determine the source of each outbreak, we build a Shortest Paths Tree (SPT) for every vertex in the corresponding component. The source is estimated as the vertex with an SPT of minimal weight.

4.2.2 MinDistB method

The method extends the MinDist approach proposed in[32], which defines the distance between viral populations as the minimum Hamming distance between their representatives. The new approach also takes into account sizes of border sets, on which the minimum distance is achieved. Formally, given an integer δ (by default $\delta = 1$), the δ-crossing between populations P_1 and P_2 is the set of pairs of variants (u, v) from different populations, the Hamming distance $D(u, v)$ between which is within δ from the minimum Hamming distance:

$$B_\delta(P_1, P_2) = \{(u, v) : u \in P_1, v \in P_2, D(u, v) \leq \min_{x \in P_1, y \in P_2} D(x, y) + \delta\}$$

(see Figure 4.3). Our empirical study shows that in case when the crossing is large (see Figure 4.3(A)), then the populations are less likely to be related than in case when the borders are small (see Figure 4.3(B)).

This effect can be intuitively explained. Two related populations likely diverge away from the common ancestor and from each other, and their borders are formed by few old survived variants closest to the common ancestor. Two unrelated populations diverging from two different ancestors may in time reduce minimum distance from each other randomly and closest variants are relatively young and abundant (see Figure 4.4).

We define a δ-distance between populations P_1 and P_2 as follows:

$$D_\delta(P_1, P_2) = D(P_1, P_2) + c \ln(|B_\delta(P_1, P_2)|) \tag{4.4}$$

where $c = 3$ is an empirically chosen constant.

4.2.2.1 Identification of genetic relatedness, transmission clusters and sources of outbreaks

For MinDistB methods, genetic relatedness graph $G = (V, E, w)$ is a weighted undirected graph with the vertex set $V = \mathcal{P}$ and an edge of weight $w_{i,j}$ connecting populations P_i, P_j whenever $w_{i,j} = D_\delta(P_1, P_2)$ does not exceed a threshold. Transmission clusters are estimated as connected components of the graph G. For each transmission cluster its source could be inferred either as a vertex with maximum eigenvector centrality or as a vertex with the shortest paths tree of minimal weight.

4.3 Results

VOICE and MinDistB were validated using experimental outbreak sequencing data, and their predictions were compared with the ReD[70] and the previously published MinDist method[32]

4.3.1 Data sets

We used the benchmark data presented in[32], which is a collection of HCV intra-host populations sampled from 335 infected individuals.

- Outbreak collection contains 142 HCV samples from 33 epidemiologically curated outbreaks reported to Centers for Disease Control and Prevention in 2008-2013. Outbreaks contain from 2 to 19 samples. Epidemiological histories, including sources of infection, are known for 10 outbreaks.
- Collection of 193 epidemiologically unrelated HCV samples.

All viral sequences represent a fragment of E1/E2 genomic region of length 264bp.

4.3.2 Prediction of epidemiological characteristics

The proposed methods were used to infer the following epidemiological characteristics:

- genetic relatedness between populations;
- transmission clusters representing outbreaks and isolated samples;
- sources of outbreaks;
- transmission directions between pairs of samples.

Comparison results are collected in Table 4.1.The variants of VOICE with deterministic and subsampling normalizations are referred to as $VOICE-D$ and $VOICE-S$, and for them we used the normalization constants $q=10$ and $q=4$, respectively. For all VOICE runs, five independent

simulations were performed, and the averages over that simulations are reported. For each simulation, VOICE-S performs 50 subsamplings, and the results of the algorithm are averaged over all subsamplings. For MinDist, sources of outbreaks were identified as vertices with highest eigenvector centralities in the corresponding genetic relatedness graphs, since for MinDist this method outperform the shortest path tree-based approach.

4.3.2.1 Genetic relatedness between populations

Viral populations from two samples are genetically related if they belong to the same outbreak and unrelated, otherwise. The genetic relatedness is validated on the union of both collections containing all outbreaks and unrelated samples. There are 55945 pairs of samples, and 479 of them are related. For all algorithms we choose the best thresholds, which produce no false positives, i.e. no unrelated populations are predicted to be related. The values of thresholds T are: $ReD : T = 2$; $MinDist : T = 11$; $MinDistB : T = 28.4$; $Voice - D : T = 1710$; $Voice - S : T = 4585$. For each method, the sensitivity (i.e. the percentage of detected related pairs) was calculated (Table 4.1). The highest sensitivity is achieved by MinDistB method. Figure 4.5 depict ROC curve for the tested methods (ReD is not present, since for this method only few viable discrete thresholds are possible). $MinDistB$ and $VOICE - D$ have highest areas under a curve value followed by $MinDist$ and $VOICE - S$.

4.3.2.2 Detection of transmission clusters

The similarities between true and estimated partitions into transmission clusters were measured using an editing metric[44], which is defined as the minimum number of elementary operations re-

quired to transform one partition into another. An elementary operation is either merging (joining of two clusters into a single cluster) or division (partition of a cluster into two clusters)[44]. We calculate sensitivity by normalizing an editing distance E by dividing it by the number N of elementary operations required to transform trivial partition (i.e. the partition into singleton sets) into the true partition. The number N is equal to $n - k$, where n is the total number of samples and k is the number of true clusters:

$$Sensitivity = \frac{E}{n-k} \times 100\%. \tag{4.5}$$

Table 4.1 shows that MinDistB and MinDist demonstrate the highest sensitivity.

4.3.2.3 Source identification

The accuracy of the source identification is defined as the percentage of correctly predicted sources for outbreaks, where the correct sources are known. The Source section of Table 4.1 shows that the best results are achieved by ReD and $VOICE - S$ which were able to detect sources in 90% of cases. At the same time, MinDist and MinDistB, which are not able to identify transmission directions, were significantly less accurate.

4.3.2.4 Transmission direction

Among tested algorithms, only ReD and $VOICE$ allows for detection of transmission directions. For that algorithms, percentages of correctly predicted pairs source-recipient were calculated (Table 4.1). Here the highest accuracy of 87.1% was achieved by ReD and $VOICE - S$.

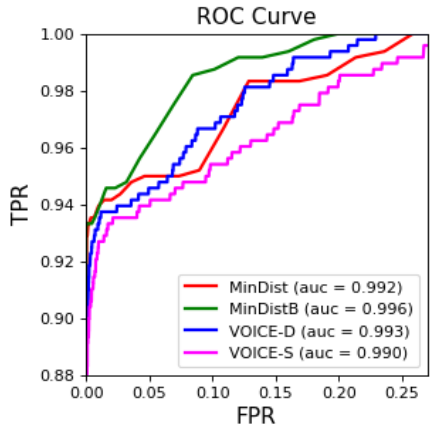

Figure 4.5 ROC curve for pairs relatedness detection

Table 4.1 Validation results

Methods	MinDist	MinDistB	ReD	VOICE-D	VOICE-S
Relatedness					
Sensitivity, %	90%	**92.9%**	55.3%	85.2%	86.8 %
AUROC	0.992	**0.996**	N/A	0.993	0.990
Clustering					
Sensitivity, %	**100%**	**100%**	96.3%	98.2%	98.2%
Source					
Accuracy, %	50%	40%	**90%**	80%	**90%**
Directions					
Accuracy, %	N/A	N/A	**87.1%**	83.9%	**87.1%**

4.3.2.5 Running time

All tests were performed on PC with DDR3-1333MHz 4GBx12 RAM and 2 Intel Xeon-X5550 2.67GHz processors. The fastest algorithms were MinDist and MinDistB, with running times 9 ms for a pair of samples in our dataset. ReD requires $\sim 0.1s$ per pair of samples, While the running time of $VOICE$ is $\sim 35s$ per pair.

4.4 Conclusions

Currently, a molecular viral analysis is one of the major approaches used for investigations of outbreaks and inference of transmission networks. Although modern sequencing technologies significantly facilitated molecular analysis, providing unprecedented access to intra-host viral populations, they generated novel bioinformatics challenges.

This work proposed two algorithms for the investigation of viral transmissions based on analysis of the intra-host viral populations, which allow clustering genetically related samples, infer transmission directions and predict sources of outbreaks. Evaluation of the algorithms on experimental data from HCV outbreaks demonstrated their ability to accurately reconstruct various transmission characteristics. However, the advantage of this method over other methods is its non-parametricity (i.e. independence from virus-specific and genomic region-specific thresholds), which makes it more universally applicable and extremely useful in situations, when the lack of training data does not allow to establish reliable relatedness thresholds.

REFERENCES

1. L. Abeler-Dörner, M. K. Grabowski, A. Rambaut, D. Pillay, C. Fraser, and PANGEA consortium. PANGEA-HIV 2: Phylogenetics and networks for generalised epidemics in africa. *Curr. Opin. HIV AIDS*, 14(3):173–180, May 2019.

2. S. Ahn and H. Vikalo. aBayesQR: A bayesian method for reconstruction of viral populations characterized by low diversity. *J. Comput. Biol.*, 25(7):637–648, July 2018.

3. M. J. Akiyama, D. Lipsey, L. Ganova-Raeva, L. Punkova, L. Agyemang, A. Sue, S. Ramachandran, Y. Khudyakov, and A. H. Litwin. A phylogenetic analysis of HCV transmission, relapse, and reinfection among people who inject drugs receiving opioid agonist therapy, 2020.

4. E. K. Alidjinou, J. Deldalle, C. Hallaert, O. Robineau, F. Ajana, P. Choisy, D. Hober, and L. Bocket. RNA and DNA sanger sequencing versus next-generation sequencing for HIV-1 drug resistance testing in treatment-naive patients. *J. Antimicrob. Chemother.*, 72(10):2823–2830, Oct. 2017.

5. S. Alroy-Preis, E. R. Daly, C. Adamski, J. Dionne-Odom, E. A. Talbot, F. Gao, S. J. Cavallo, K. Hansen, J. C. Mahoney, E. Metcalf, C. Loring, C. Bean, J. Drobeniuc, G.-L. Xia, S. Kamili, J. T. Montero, and New Hampshire and Centers for Disease Control and Prevention Investigation Teams. Large outbreak of hepatitis C virus associated with drug diversion by a healthcare technician. *Clin. Infect. Dis.*, 67(6):845–853, Aug. 2018.

6. A. Apostolou, M. L. Bartholomew, R. Greeley, S. M. Guilfoyle, M. Gordon, C. Genese, J. P.

Davis, B. Montana, and G. Borlaug. Transmission of hepatitis c virus associated with surgical procedures-new jersey 2010 and wisconsin 2011. *MMWR. Morbidity and mortality weekly report*, 64(7):165–170, 2015.

7. J. Archer, M. S. Braverman, B. E. Taillon, B. Desany, I. James, P. R. Harrigan, M. Lewis, and D. L. Robertson. Detection of low-frequency pretherapy chemokine (cxc motif) receptor 4-using hiv-1 with ultra-deep pyrosequencing. *AIDS (London, England)*, 23(10):1209, 2009.
8. A. Arias, P. López, R. Sánchez, Y. Yamamura, and V. Rivera-Amill. Sanger and next generation sequencing approaches to evaluate HIV-1 virus in blood compartments. *Int. J. Environ. Res. Public Health*, 15(8), Aug. 2018.
9. A. Artyomenko, N. C. Wu, S. Mangul, E. Eskin, R. Sun, and A. Zelikovsky. Long Single-Molecule reads can resolve the complexity of the influenza virus composed of rare, closely related mutant variants. *J. Comput. Biol.*, 24(6):558–570, June 2017.
10. Y. Assefa and C. F. Gilks. Second-line antiretroviral therapy: so much to be done. *Lancet HIV*, 4(10):e424–e425, Oct. 2017.
11. I. V. Astrakhantseva, D. S. Campo, A. Araujo, C.-G. Teo, Y. Khudyakov, and S. Kamili. Differences in variability of hypervariable region 1 of hepatitis C virus (HCV) between acute and chronic stages of HCV infection. *In Silico Biol.*, 11(5-6):163–173, 2011.
12. I. Astrovskaya, B. Tork, S. Mangul, K. Westbrooks, I. Măndoiu, P. Balfe, and A. Zelikovsky. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC Bioinformatics*, 12 Suppl 6:S1, July 2011.
13. J. A. Baaijens, A. Z. El Aabidine, E. Rivals, and A. Schönhuth. De novo assembly of viral

quasispecies using overlap graphs, 2017.

14. S. Barik, S. Das, and H. Vikalo. Viral quasispecies reconstruction via correlation clustering. *bioRxiv*, page 096768, 2016.

15. S. Basodi, P. B. Icer, P. Skums, Y. Khudyakov, A. Zelikovsky, and Y. Pan. Classification of HCV infections through sequence image normalization, 2017.

16. P. I. Baykal, A. Artyomenko, S. Ramachandran, Y. Khudyakov, A. Zelikovsky, and P. Skums. Assessment of HCV infection stage as recent or chronic using multi-parameter analysis and machine learning. In *2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 1–1, Oct. 2017.

17. N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.*, 31(13):3850–3855, July 2003.

18. N. Beerenwinkel, T. Sing, T. Lengauer, J. Rahnenführer, K. Roomp, I. Savenkov, R. Fischer, D. Hoffmann, J. Selbig, K. Korn, H. Walter, T. Berg, P. Braun, G. Fätkenheuer, M. Oette, J. Rockstroh, B. Kupfer, R. Kaiser, and M. Däumer. Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21(21):3943–3950, Nov. 2005.

19. P. Bellecave, P. Recordon-Pinson, J. Papuchon, M.-A. Vandenhende, S. Reigadas, B. Tauzin, and H. Fleury. Detection of low-frequency HIV type 1 reverse transcriptase drug resistance mutations by ultradeep sequencing in naive HIV type 1-infected individuals. *AIDS Res. Hum. Retroviruses*, 30(2):170–173, Feb. 2014.

20. S. Benidt and D. Nettleton. Simseq: a nonparametric approach to simulation of rna-sequence datasets. *Bioinformatics*, 31(13):2131–2140, 2015.

21. C. Beyrer and A. Pozniak. HIV drug resistance — an emerging threat to epidemic control, 2017.

22. C. Bleidorn. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *System. Biodivers.*, 14(1):1–8, Jan. 2016.

23. V. F. Boltz, J. Rausch, W. Shao, J. Hattori, B. Luke, F. Maldarelli, J. W. Mellors, M. F. Kearney, and J. M. Coffin. Ultrasensitive single-genome sequencing: accurate, targeted, next generation sequencing of HIV-1 RNA. *Retrovirology*, 13(1):87, Dec. 2016.

24. A. C. Bourgeois, M. Edmunds, A. Awan, L. Jonah, O. Varsaneux, and W. Siu. HIV in Canada-Surveillance report, 2016. *Can. Commun. Dis. Rep.*, 43(12):248–256, Dec. 2017.

25. C. Bron and J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, Sept. 1973. ISSN 0001-0782. doi: 10.1145/362342.362367. URL `http://doi.acm.org/10.1145/362342.362367`.

26. E. M. Campbell, H. Jia, A. Shankar, D. Hanson, W. Luo, S. Masciotra, S. M. Owen, A. M. Oster, R. R. Galang, M. W. Spiller, S. J. Blosser, E. Chapman, J. C. Roseberry, J. Gentry, P. Pontones, J. Duwve, P. Peyrani, R. M. Kagan, J. M. Whitcomb, P. J. Peters, W. Heneine, J. T. Brooks, and W. M. Switzer. Detailed transmission network analysis of a large Opiate-Driven outbreak of HIV infection in the united states. *J. Infect. Dis.*, 216(9):1053–1062, Nov. 2017.

27. D. S. Campo, Z. Dimitrova, L. Yamasaki, P. Skums, D. T. Lau, G. Vaughan, J. C. Forbi,

C.-G. Teo, and Y. Khudyakov. Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC Genomics*, 15 Suppl 5:S4, July 2014.

28. D. S. Campo, Z. Dimitrova, L. Yamasaki, P. Skums, D. T. Lau, G. Vaughan, J. C. Forbi, C.-G. Teo, and Y. Khudyakov. Next-generation sequencing reveals large connected networks of intra-host hcv variants. *BMC genomics*, 15(Suppl 5):S4, 2014.
29. D. S. Campo, P. Skums, Z. Dimitrova, G. Vaughan, J. C. Forbi, C. G. Teo, Y. Khudyakov, and D. T.-Y. Lau. Drug resistance of a viral population and its individual intrahost variants during the first 48 hours of therapy. *Clin. Pharmacol. Ther.*, 95(6):627–635, June 2014.
30. D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, S. Sims, I. Rytsareva, G. Vaughan, H.-J. Roh, M. A. Purdy, A. Sue, and Y. Khudyakov. Accurate genetic detection of hepatitis C virus transmissions in outbreak settings. *J. Infect. Dis.*, 213(6):957–965, Mar. 2016.
31. D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *Journal of Infectious Diseases*, 213(6):957–965, 2016.
32. D. S. Campo, G.-L. Xia, Z. Dimitrova, Y. Lin, J. C. Forbi, L. Ganova-Raeva, L. Punkova, S. Ramachandran, H. Thai, P. Skums, et al. Accurate genetic detection of hepatitis c virus transmissions in outbreak settings. *Journal of Infectious Diseases*, 213(6):957–965, 2016.
33. M. R. Capobianchi, E. Giombini, and G. Rozera. Next-generation sequencing technology in clinical virology, 2013.
34. L. A. Carlisle, T. Turk, K. Kusejko, K. J. Metzner, C. Leemann, C. D. Schenkel, N. Bach-

mann, S. Posada, N. Beerenwinkel, J. Böni, S. Yerly, T. Klimkait, M. Perreau, D. L. Braun, A. Rauch, A. Calmy, M. Cavassini, M. Battegay, P. Vernazza, E. Bernasconi, H. F. Günthard, R. D. Kouyos, and Swiss HIV Cohort Study. Viral diversity based on Next-Generation sequencing of HIV-1 provides precise estimates of infection recency and time since infection. *J. Infect. Dis.*, 220(2):254–265, June 2019.

35. K. Cashin, L. R. Gray, K. L. Harvey, D. Perez-Bercoff, G. Q. Lee, J. Sterjovski, M. Roche, J. F. Demarest, F. Drummond, P. Richard Harrigan, M. J. Churchill, and P. R. Gorry. Reliable genotypic tropism tests for the major HIV-1 subtypes, 2015.

36. S. B. Chabria, S. Gupta, and M. J. Kozal. Deep sequencing of HIV: clinical and research applications. *Annu. Rev. Genomics Hum. Genet.*, 15:295–325, May 2014.

37. J. Chen, Y. Zhao, and Y. Sun. De novo haplotype reconstruction in viral quasispecies using paired-end read guided path finding, 2018.

38. M. Cornelissen, A. Gall, M. Vink, F. Zorgdrager, Š. Binter, S. Edwards, S. Jurriaans, M. Bakker, S. H. Ong, L. Gras, A. van Sighem, D. Bezemer, F. de Wolf, P. Reiss, P. Kellam, B. Berkhout, C. Fraser, A. C. van der Kuyl, and BEEHIVE Consortium. From clinical sample to complete genome: Comparing methods for the extraction of HIV-1 RNA for high-throughput deep sequencing. *Virus Res.*, 239:10–16, July 2017.

39. M. Cruz-Rivera, J. C. Forbi, L. H. T. Yamasaki, C. A. Vazquez-Chacon, A. Martinez-Guarneros, J. C. Carpio-Pedroza, A. Escobar-Gutiérrez, K. Ruiz-Tovar, S. Fonseca-Coronado, and G. Vaughan. Molecular epidemiology of viral diseases in the era of next generation sequencing. *J. Clin. Virol.*, 57(4):378–380, Aug. 2013.

40. N. De Maio, C.-H. Wu, and D. J. Wilson. SCOTTI: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.*, 12(9):e1005130, Sept. 2016.

41. N. De Maio, C.-H. Wu, and D. J. Wilson. Scotti: Efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS computational biology*, 12(9): e1005130, 2016.

42. A. Derache, C. C. Iwuji, K. Baisley, S. Danaviah, A.-G. Marcelin, V. Calvez, T. de Oliveira, F. Dabis, K. Porter, and D. Pillay. Impact of next-generation sequencing defined human immunodeficiency virus pretreatment drug resistance on virological outcomes in the ANRS 12249 Treatment-as-Prevention trial, 2019.

43. A. Derache, C. C. Iwuji, S. Danaviah, J. Giandhari, A.-G. Marcelin, V. Calvez, T. de Oliveira, F. Dabis, D. Pillay, and R. K. Gupta. Predicted antiviral activity of tenofovir versus abacavir in combination with a cytosine analogue and the integrase inhibitor dolutegravir in HIV-1-infected south african patients initiating or failing first-line ART, 2019.

44. M. M. Deza and E. Deza. Encyclopedia of distances, 2009.

45. E. Domingo and J. J. Holland. RNA VIRUS MUTATIONS AND FITNESS FOR SURVIVAL, 1997.

46. E. Domingo, E. Martínez-Salas, F. Sobrino, J. C. de la Torre, A. Portela, J. Ortín, C. López-Galindez, P. Pérez-Breña, N. Villanueva, R. Nájera, S. VandePol, D. Steinhauer, N. DePolo, and J. Holland. The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance — a review, 1985.

47. E. Domingo, J. Sheldon, and C. Perales. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.*, 76(2):159–216, June 2012.

48. P. Domingo-Calap, E. Segredo-Otero, M. Durán-Moreno, and R. Sanjuán. Social evolution of innate immunity evasion in a virus. *Nat Microbiol*, 4(6):1006–1013, June 2019.

49. M. Döring, J. Büch, G. Friedrich, A. Pironti, P. Kalaghatgi, E. Knops, E. Heger, M. Obermeier, M. Däumer, A. Thielen, R. Kaiser, T. Lengauer, and N. Pfeifer. geno2pheno[ngs-freq]: a genotypic interpretation system for identifying viral drug resistance using next-generation sequencing data. *Nucleic Acids Res.*, 46(W1):W271–W277, July 2018.

50. D. C. Douek, P. D. Kwong, and G. J. Nabel. The rational design of an AIDS vaccine, 2006.

51. J. W. Drake and J. J. Holland. Mutation rates among RNA viruses, 1999.

52. M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species, 1988.

53. M. Eigen, J. McCaskill, and P. Schuster. The molecular quasi-species. *Advances in chemical physics*, 75:149–263, 1989.

54. A. Eliseev, K. M. Gibson, P. Avdeyev, D. Novik, M. L. Bendall, M. Pérez-Losada, N. Alexeev, and K. A. Crandall. Evaluation of haplotype callers for next-generation sequencing of viruses. Nov. 2019.

55. N. Eriksson, L. Pachter, Y. Mitsuya, S.-Y. Rhee, C. Wang, B. Gharizadeh, M. Ronaghi, R. W. Shafer, and N. Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, 4(4):e1000074, May 2008.

56. A. F. Feder, S.-Y. Rhee, S. P. Holmes, R. W. Shafer, D. A. Petrov, and P. S. Pennings. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1.

Elife, 5, Feb. 2016.

57. L. Ferretti, C. Tennakoon, A. Silesian, and G. F. A. Ribeca. SiNPle: Fast and sensitive variant calling for deep sequencing data. *Genes*, 10(8), July 2019.

58. N. Fischer, D. Indenbirken, T. Meyer, M. Lütgehetmann, H. Lellek, M. Spohn, M. Aepfelbacher, M. Alawi, and A. Grundhoff. Evaluation of unbiased Next-Generation sequencing of RNA (RNA-seq) as a diagnostic method in influenza Virus-Positive respiratory samples. *J. Clin. Microbiol.*, 53(7):2238–2250, July 2015.

59. W. Fischer, V. V. Ganusov, E. E. Giorgi, P. T. Hraber, B. F. Keele, T. Leitner, C. S. Han, C. D. Gleasner, L. Green, C.-C. Lo, A. Nag, T. C. Wallstrom, S. Wang, A. J. McMichael, B. F. Haynes, B. H. Hahn, A. S. Perelson, P. Borrow, G. M. Shaw, T. Bhattacharya, and B. T. Korber. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One*, 5(8):e12303, Aug. 2010.

60. R. G. Fisher, D. M. Smith, B. Murrell, R. Slabbert, B. M. Kirby, C. Edson, M. F. Cotton, R. H. Haubrich, S. L. Kosakovsky Pond, and G. U. Van Zyl. Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure. *J. Clin. Virol.*, 62:48–53, Jan. 2015.

61. P. Flaherty, G. Natsoulis, O. Muralidharan, M. Winters, J. Buenrostro, J. Bell, S. Brown, M. Holodniy, N. Zhang, and H. P. Ji. Ultrasensitive detection of rare mutations using next-generation targeted resequencing. *Nucleic Acids Res.*, 40(1):e2, Jan 2012.

62. W. F. Flynn, M. W. Chang, Z. Tan, G. Oliveira, J. Yuan, J. F. Okulicz, B. E. Torbett, and R. M. Levy. Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of

correlated mutations in gag and protease. *PLoS Comput. Biol.*, 11(4):e1004249, Apr. 2015.

63. J. Fokam, M. C. Bellocchi, D. Armenia, A. J. Nanfack, L. Carioti, F. Continenza, D. Takou, E. S. Temgoua, C. Tangimpundu, J. N. Torimiro, P. N. Koki, C. N. Fokunang, G. Cappelli, A. Ndjolo, V. Colizzi, F. Ceccherini-Silberstein, C.-F. Perno, and M. M. Santoro. Next-generation sequencing provides an added value in determining drug resistance and viral tropism in cameroonian HIV-1 vertically infected children. *Medicine*, 97(13):e0176, Mar. 2018.

64. S. Fourati and J.-M. Pawlotsky. Virologic tools for HCV drug resistance testing. *Viruses*, 7 (12):6346–6359, Dec. 2015.

65. B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354–2360, June 2002.

66. E. Gerasimov. *Analysis of NGS Data from Immune Response and Viral Samples*. PhD thesis, Georgia State University, 2017.

67. F. D. Giallonardo, A. Töpfer, M. Rey, S. Prabhakaran, Y. Duport, C. Leemann, S. Schmutz, N. K. Campbell, B. Joos, M. R. Lecca, A. Patrignani, M. Däumer, C. Beisel, P. Rusert, A. Trkola, H. F. Günthard, V. Roth, N. Beerenwinkel, and K. J. Metzner. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Research*, 42(14):e115, 2014. doi: 10.1093/nar/gku537. URL `+http://dx.doi.org/10.1093/nar/gku537`.

68. K. M. Gibson, M. C. Steiner, S. Kassaye, F. Maldarelli, Z. Grossman, M. Pérez-Losada, and

K. A. Crandall. Corrigendum: A 28-year history of HIV-1 drug resistance and transmission in washington, DC, 2019.

69. M. I. Gismondi, J. M. Díaz Carrasco, P. Valva, P. D. Becker, C. A. Guzmán, R. H. Campos, and M. V. Preciado. Dynamic changes in viral population structure and compartmentalization during chronic hepatitis C virus infection in children. *Virology*, 447(1-2):187–196, Dec. 2013.

70. O. Glebova, S. Knyazev, A. Melnyk, A. Artyomenko, Y. Khudyakov, A. Zelikovsky, and P. Skums. Inference of genetic relatedness between viral quasispecies from sequencing data. *BMC Genomics*, 18(Suppl 10):918, Dec. 2017.

71. M. Gwinn, D. MacCannell, and G. L. Armstrong. Next-Generation sequencing of infectious pathogens, 2019.

72. J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, Dec. 2018.

73. B. Hajarizadeh, J. Grebely, and G. J. Dore. Epidemiology and natural history of HCV infection. *Nat. Rev. Gastroenterol. Hepatol.*, 10(9):553–562, Sept. 2013.

74. R. L. Hamers and R. Paredes. Next-generation sequencing and HIV drug resistance surveillance, 2016.

75. O. Harismendy, R. B. Schwab, L. Bao, J. Olson, S. Rozenzhak, S. K. Kotsopoulos, S. Pond, B. Crain, M. S. Chee, K. Messer, D. R. Link, and K. A. Frazer. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome Biol.*, 12

(12):R124, 2011.

76. M. R. Henn, C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, A. R. Macalalad, A. M. Berlin, C. M. Malboeuf, E. M. Ryan, S. Gnerre, M. C. Zody, R. L. Erlich, L. M. Green, A. Berical, Y. Wang, M. Casali, H. Streeck, A. K. Bloom, T. Dudek, D. Tully, R. Newman, K. L. Axten, A. D. Gladden, L. Battis, M. Kemper, Q. Zeng, T. P. Shea, S. Gujja, C. Zedlack, O. Gasser, C. Brander, C. Hess, H. F. Günthard, Z. L. Brumme, C. J. Brumme, S. Bazner, J. Rychert, J. P. Tinsley, K. H. Mayer, E. Rosenberg, F. Pereyra, J. Z. Levin, S. K. Young, H. Jessen, M. Altfeld, B. W. Birren, B. D. Walker, and T. M. Allen. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog.*, 8(3):e1002529, Mar. 2012.

77. T. Hoenen, A. Groseth, K. Rosenke, R. J. Fischer, A. Hoenen, S. D. Judson, C. Martellaro, D. Falzarano, A. Marzi, R. B. Squires, K. R. Wollenberg, E. de Wit, J. Prescott, D. Safronetz, N. van Doremalen, T. Bushmaker, F. Feldmann, K. McNally, F. K. Bolay, B. Fields, T. Sealy, M. Rayfield, S. T. Nichol, K. C. Zoon, M. Massaquoi, V. J. Munster, and H. Feldmann. Nanopore sequencing as a rapidly deployable ebola outbreak tool. *Emerg. Infect. Dis.*, 22(2): 331–334, Feb. 2016.

78. C. Hoffmann, N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas, and F. D. Bushman. Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations. *Nucleic acids research*, 35(13):e91, 2007.

79. J. J. Holland, J. C. De La Torre, and D. A. Steinhauer. RNA virus populations as quasispecies, 1992.

80. L. Z. Hong, S. Hong, H. T. Wong, P. P. K. Aw, Y. Cheng, A. Wilm, P. F. de Sessions, S. G. Lim, N. Nagarajan, M. L. Hibberd, S. R. Quake, and W. F. Burkholder. BAsE-Seq: a method for obtaining long viral haplotypes from short sequence reads, 2014.

81. A. Huang, R. Kantor, A. DeLong, L. Schreier, and S. Istrail. QColors: An algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads, 2011.

82. M. Huber, K. J. Metzner, F. D. Geissberger, C. Shah, C. Leemann, T. Klimkait, J. Böni, A. Trkola, and O. Zagordi. MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *J. Virol. Methods*, 240:7–13, Feb. 2017.

83. M. Hunt, A. Gall, S. H. Ong, J. Brener, B. Ferns, P. Goulder, E. Nastouli, J. A. Keane, P. Kellam, and T. D. Otto. IVA: accurate de novo assembly of RNA virus genomes. *Bioinformatics*, 31(14):2374–2376, July 2015.

84. K. K. Irwin, N. Renzette, T. F. Kowalik, and J. D. Jensen. Antiviral drug resistance as an adaptive process. *Virus Evol*, 2(1):vew014, Jan. 2016.

85. O. Isakov, A. V. Bordería, D. Golan, A. Hamenahem, G. Celniker, L. Yoffe, H. Blanc, M. Vignuzzi, and N. Shomron. Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics*, 31(13):2141–2150, July 2015.

86. K. Jair, C. D. McCann, H. Reed, A. D. Castel, M. Pérez-Losada, B. Wilbourn, A. E. Greenberg, J. A. Jordan, and the DC Cohort Executive Committee. Validation of publicly-available software used in analyzing NGS data for HIV-1 drug resistance mutations and transmission

networks in a washington, DC, cohort, 2019.

87. D. Jayasundara, I. Saeed, S. Maheswararajah, B. C. Chang, S.-L. Tang, and S. K. Halgamuge. ViQuaS: an improved reconstruction pipeline for viral quasispecies spectra generated by next-generation sequencing. *Bioinformatics*, 31(6):886–896, Mar. 2015.

88. J. A. Johnson, J.-F. Li, X. Wei, J. Lipscomb, D. Irlbeck, C. Craig, A. Smith, D. E. Bennett, M. Monsour, P. Sandstrom, E. Randall Lanier, and W. Heneine. Minority HIV-1 drug resistance mutations are present in antiretroviral Treatment–Naïve populations and associate with reduced treatment efficacy. *PLoS Med.*, 5(7):e158, July 2008.

89. V. Jojic, T. Hertz, and N. Jojic. Population sequencing using short reads: HIV as a case study. *Pac. Symp. Biocomput.*, pages 114–125, 2008.

90. T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach, 2011.

91. T. Jombart, A. Cori, X. Didelot, S. Cauchemez, C. Fraser, and N. Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.*, 10(1):e1003457, Jan. 2014.

92. P. H. Kilmarx. Global epidemiology of HIV. *Curr. Opin. HIV AIDS*, 4(4):240–246, July 2009.

93. D. E. Kireev, A. E. Lopatukhin, A. V. Murzakova, E. V. Pimkina, A. S. Speranskaya, A. D. Neverov, G. G. Fedonin, Y. S. Fantin, and G. A. Shipulin. Evaluating the accuracy and sensitivity of detecting minority HIV-1 populations by Illumina next-generation sequencing. *J. Virol. Methods*, 261:40–45, 11 2018.

94. D. Klinkenberg, J. A. Backer, X. Didelot, C. Colijn, and J. Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.*, 13(5):e1005495, May 2017.

95. S. Knyazev, V. Tsyvina, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, and A. Zelikovsky. CliqueSNV: Scalable reconstruction of intra-host viral populations from ngs reads. biorxiv. *biorxiv*, 2018.

96. D. C. Koboldt, K. Chen, T. Wylie, D. E. Larson, M. D. McLellan, E. R. Mardis, G. M. Weinstock, R. K. Wilson, and L. Ding. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285, Sept. 2009.

97. S. L. Kosakovsky Pond, S. Weaver, A. J. Leigh Brown, and J. O. Wertheim. HIV-TRACE (TRAnsmission cluster engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. *Mol. Biol. Evol.*, 35(7):1812–1819, July 2018.

98. J. Kováč. Complexity of the path avoiding forbidden pairs problem revisited. *Discrete Appl. Math.*, 161(10-11):1506–1512, July 2013. ISSN 0166-218X. doi: 10.1016/j.dam.2012.12.022. URL `http://dx.doi.org/10.1016/j.dam.2012.12.022`.

99. C. Kuiken, B. Korber, and R. W. Shafer. HIV sequence databases. *AIDS Rev.*, 5(1):52–61, Jan. 2003.

100. B. A. Larder, A. Kohli, P. Kellam, S. D. Kemp, M. Kronick, and R. D. Henfrey. Quantitative detection of HIV-1 drug resistance mutations by automated DNA sequencing. *Nature*, 365 (6447):671–673, Oct. 1993.

101. C. Latkin, C. Yang, A. K. Srikrishnan, S. Solomon, S. H. Mehta, D. D. Celentano, M. S.

Kumar, A. Knowlton, and S. S. Solomon. The relationship between social network factors, HIV, and hepatitis C among injection drug users in chennai, india. *Drug Alcohol Depend.*, 117(1):50–54, Aug. 2011.

102. E. Levina and P. Bickel. The earthmover's distance is the mallows distance: Some insights from statistics. *Proceedings of ICCV 2001*, pages 251–256, 2001.

103. S. Leviyang, I. Griva, S. Ita, and W. E. Johnson. A penalized regression approach to haplotype reconstruction of viral populations arising in early HIV/SIV infection. *Bioinformatics*, 33 (16):2455–2463, Aug. 2017.

104. T. F. Liu and R. W. Shafer. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect. Dis.*, 42(11):1608–1618, June 2006.

105. A. G. Longmire, S. Sims, I. Rytsareva, D. S. Campo, P. Skums, Z. Dimitrova, S. Ramachandran, M. Medrzycki, H. Thai, L. Ganova-Raeva, Y. Lin, L. T. Punkova, A. Sue, M. Mirabito, S. Wang, R. Tracy, V. Bolet, T. Sukalac, C. Lynberg, and Y. Khudyakov. GHOST: global hepatitis outbreak and surveillance technology. *BMC Genomics*, 18(Suppl 10):916, Dec. 2017.

106. E. Lontok, P. Harrington, A. Howe, T. Kieffer, J. Lennerstrand, O. Lenz, F. McPhee, H. Mo, N. Parkin, T. Pilot-Matias, and V. Miller. Hepatitis C virus drug resistance-associated substitutions: State of the art summary. *Hepatology*, 62(5):1623–1632, Nov. 2015.

107. R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study

2010. *The lancet*, 380(9859):2095–2128, 2012.

108. A. R. Macalalad, M. C. Zody, P. Charlebois, N. J. Lennon, R. M. Newman, C. M. Malboeuf, E. M. Ryan, C. L. Boutwell, K. A. Power, D. E. Brackney, K. N. Pesko, J. Z. Levin, G. D. Ebel, T. M. Allen, B. W. Birren, and M. R. Henn. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput. Biol.*, 8(3):e1002417, Mar. 2012.

109. R. Malhotra, M. M. S. Wu, A. Rodrigo, M. Poss, and R. Acharya. Maximum likelihood de novo reconstruction of viral populations using paired end sequencing data. Feb. 2015.

110. R. Malhotra, M. Jha, M. Poss, and R. Acharya. A random forest classifier for detecting rare variants in NGS data from viral populations. *Comput. Struct. Biotechnol. J.*, 15:388–395, July 2017.

111. C. L. Mallows. A note on asymptotic joint normality. *Annals of Mathematical Statistics*, 43 (2):508–515, 1972.

112. N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Măndoiu, and A. Zelikovsky. Reconstructing viral quasispecies from NGS amplicon reads. *In Silico Biol.*, 11(5-6):237–249, 2011.

113. I. Mandoiu and A. Zelikovsky. *Computational Methods for Next Generation Sequencing Data Analysis*. John Wiley & Sons, Sept. 2016.

114. S. Mangul, N. C. Wu, N. Mancuso, A. Zelikovsky, R. Sun, and E. Eskin. Accurate viral population assembly from ultra-deep sequencing data, 2014.

115. M. Martell, J. I. Esteban, J. Quer, J. Genescà, A. Weiner, R. Esteban, J. Guardia, and

J. Gómez. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J. Virol.*, 66(5):3225–3229, May 1992.

116. J. McGinnis, J. Laplante, M. Shudt, and K. St George. Corrigendum to "next generation sequencing for whole genome analysis and surveillance of influenza a viruses" [j. clin. virol. 79 (2016) 44-50]. *J. Clin. Virol.*, 93:65, Aug. 2017.

117. K. S. McKeegan, M. I. Borges-Walmsley, and A. R. Walmsley. Microbial and viral drug resistance mechanisms. *Trends Microbiol.*, 10(10 Suppl):S8–14, 2002.

118. A. Melnyk, S. Knyazev, Y. Khudyakov, F. Vannberg, L. Bunimovich, P. Skums, and A. Zelikovsky. Using earth mover's distance for viral outbreak investigations. *bioRxiv*, page 628859, 2019.

119. K. J. Metzner, P. Rauch, H. Walter, C. Boesecke, B. Zöllner, H. Jessen, K. Schewe, S. Fenske, H. Gellermann, and H.-J. Stellbrink. Detection of minor populations of drug-resistant HIV-1 in acute seroconverters. *AIDS*, 19(16):1819–1825, Nov. 2005.

120. K. Mitchell, I. Mandric, J. Brito, Q. Wu, S. Knyazev, S. Chang, L. S. Martin, A. Karlsberg, E. Gerasimov, R. Littman, B. L. Hill, N. C. Wu, H. Yang, K. Hsieh, L. Chen, T. Shabani, G. Shabanets, D. Yao, R. Sun, J. Schroeder, E. Eskin, A. Zelikovsky, P. Skums, M. Pop, and S. Mangul. Benchmarking of computational error-correction methods for next-generation sequencing data. *biorxiv.org*, 2019.

121. N. Mollentze, L. H. Nel, S. Townsend, K. le Roux, K. Hampson, D. T. Haydon, and S. Soubeyrand. A bayesian approach for inferring the dynamics of partially observed en-

demic infectious diseases from space-time-genetic data, 2014.

122. V. Montoya, A. D. Olmstead, N. Z. Janjua, P. Tang, J. Grebely, D. Cook, P. Richard Harrigan, and M. Krajden. Differentiation of acute from chronic hepatitis C virus infection by nonstructural 5B deep sequencing: A population-level tool for incidence estimation, 2015.

123. M. J. Morelli, G. Thébaud, J. Chadœuf, D. P. King, D. T. Haydon, and S. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput. Biol.*, 8(11):e1002768, Nov. 2012.

124. M. Nicolae, S. Mangul, I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011. URL `http://www.almob.org/content/6/1/9`.

125. M. Noguera-Julian. HIV drug resistance testing – the quest for Point-of-Care, 2019.

126. M. Obermeier, A. Pironti, T. Berg, P. Braun, M. Däumer, J. Eberle, R. Ehret, R. Kaiser, N. Kleinkauf, K. Korn, C. Kücherer, H. Müller, C. Noah, M. Stürmer, A. Thielen, E. Wolf, and H. Walter. HIV-GRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology*, 55(2):102–107, Jan. 2012.

127. A. M. Oster, A. M. France, N. Panneer, M. Cheryl Bañez Ocfemia, E. Campbell, S. Dasgupta, W. M. Switzer, J. O. Wertheim, and A. L. Hernandez. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data, 2018.

128. S. D. Pawar, C. Freas, I. T. Weber, and R. W. Harrison. Analysis of drug resistance in HIV protease. *BMC Bioinformatics*, 19(Suppl 11):362, Oct. 2018.

129. P. J. Peters, P. Pontones, K. W. Hoover, M. R. Patel, R. R. Galang, J. Shields, S. J. Blosser,

M. W. Spiller, B. Combs, W. M. Switzer, C. Conrad, J. Gentry, Y. Khudyakov, D. Waterhouse, S. M. Owen, E. Chapman, J. C. Roseberry, V. McCants, P. J. Weidle, D. Broz, T. Samandari, J. Mermin, J. Walthall, J. T. Brooks, J. M. Duwve, and Indiana HIV Outbreak Investigation Team. HIV infection linked to injection use of oxymorphone in indiana, 2014-2015. *N. Engl. J. Med.*, 375(3):229–239, July 2016.

130. A. Pizzorno, Y. Abed, and G. Boivin. Influenza drug resistance. *Semin. Respir. Crit. Care Med.*, 32(4):409–422, Aug. 2011.

131. J. A. Polonsky, A. Baidjoe, Z. N. Kamvar, A. Cori, K. Durski, W. J. Edmunds, R. M. Eggo, S. Funk, L. Kaiser, P. Keating, O. l. P. de Waroux, M. Marks, P. Moraga, O. Morgan, P. Nouvellet, R. Ratnayake, C. H. Roberts, J. Whitworth, and T. Jombart. Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 374(1776):20180276, July 2019.

132. S. Posada-Cespedes, D. Seifert, and N. Beerenwinkel. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res.*, 239:17–32, July 2017.

133. S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. HIV haplotype inference using a propagating dirichlet process mixture model, 2014.

134. M. C. F. Prosperi and M. Salemi. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1):132–133, Jan. 2012.

135. M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1):341,

2012.

136. J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. Baum, B. Becker-Ziaja, J.-P. Boettcher, M. Cabeza-Cabrerizo, A. Camino-Sanchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. G. G. Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallash, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, R. Trina, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune, K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Woelfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keita, A. Rambaut, P. Formenty, S. Gunther, and M. W. Carroll. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228–232, Feb. 2016.

137. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, Jr, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simp-

son, O. G. Pybus, K. G. Andersen, and N. J. Loman. Multiplex PCR method for MinION and illumina sequencing of zika and other virus genomes directly from clinical samples. *Nat. Protoc.*, 12(6):1261–1276, June 2017.

138. A. Quirin, O. Cordón, V. P. Guerrero-Bote, B. Vargas-Quesada, and F. Moya-Anegón. A quick mst-based algorithm to obtain pathfinder networks. *Journal of the American Society for Information Science and Technology*, 59(12):1912–1924, 2008.

139. S. Ramachandran, D. S. Campo, Z. E. Dimitrova, G.-L. Xia, M. A. Purdy, and Y. E. Khudyakov. Temporal variations in the hepatitis C virus intrahost population during chronic infection. *J. Virol.*, 85(13):6369–6380, July 2011.

140. S. Ramachandran, H. Thai, J. C. Forbi, R. R. Galang, Z. Dimitrova, G.-L. Xia, Y. Lin, L. T. Punkova, P. R. Pontones, J. Gentry, S. J. Blosser, J. Lovchik, W. M. Switzer, E. Teshale, P. Peters, J. Ward, Y. Khudyakov, and Hepatitis C Investigation Team. A large HCV transmission network enabled a fast-growing HIV outbreak in rural indiana, 2015. *EBioMedicine*, 37:374–381, Nov. 2018.

141. D. A. Rasmussen, E. M. Volz, and K. Koelle. Phylodynamic inference for structured epidemiological models. *PLoS Comput. Biol.*, 10(4):e1003570, Apr. 2014.

142. O. Ratmann, M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, A. Gall, P. Kellam, D. Pillay, J. Kagaayi, G. Kigozi, T. C. Quinn, M. J. Wawer, O. Laeyendecker, D. Serwadda, R. H. Gray, C. Fraser, and PANGEA Consortium and Rakai Health Sciences Program. Inferring HIV-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylo-

genetic analysis. *Nat. Commun.*, 10(1):1411, Mar. 2019.

143. C. M. Research and Case Medical Research. FDA authorizes marketing of first next-generation sequencing test for detecting HIV-1 drug resistance mutations, 2019.

144. S.-Y. Rhee, T. F. Liu, S. P. Holmes, and R. W. Shafer. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput. Biol.*, 3(5):e87, May 2007.

145. M. Riemenschneider and D. Heider. Current approaches in computational drug resistance prediction in HIV. *Curr. HIV Res.*, 14(4):307–315, 2016.

146. F. Rodriguez-Frias, M. Buti, D. Tabernero, and M. Homs. Quasispecies structure, cornerstone of hepatitis B virus infection: mass sequencing approach. *World J. Gastroenterol.*, 19(41): 6995–7023, Nov. 2013.

147. E. O. Romero-Severson, I. Bulla, and T. Leitner. Phylogenetically resolving epidemiologic linkage. *Proc. Natl. Acad. Sci. U. S. A.*, 113(10):2690–2695, Mar. 2016.

148. P. Rosenthal. Faculty of 1000 evaluation for hepatitis C virus drug resistance-associated substitutions: State of the art summary, 2015.

149. A. Routh, M. W. Chang, J. F. Okulicz, J. E. Johnson, and B. E. Torbett. CoVaMa: Co-Variation mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. *Methods*, 91:40–47, Dec. 2015.

150. W. Rutvisuttinunt, P. Chinnawirotpisan, S. Simasathien, S. K. Shrestha, I.-K. Yoon, C. Klungthong, and S. Fernandez. Simultaneous and complete genome sequencing of influenza a and B with high coverage by illumina MiSeq platform, 2013.

151. I. Rytsareva, D. S. Campo, Y. Zheng, S. Sims, S. V. Thankachan, C. Tetik, J. Chirag, S. P.

Chockalingam, A. Sue, S. Aluru, and Y. Khudyakov. Efficient detection of viral transmissions with Next-Generation sequencing data. *BMC Genomics*, 18(Suppl 4):372, May 2017.

152. M. Salemi. The intra-host evolutionary and population dynamics of human immunodeficiency virus type 1: a phylogenetic perspective. *Infect. Dis. Rep.*, 5(Suppl 1):e3, June 2013.

153. R. W. Shafer. Rationale and uses of a public HIV Drug-Resistance database. *J. Infect. Dis.*, 194(Supplement_1):S51–S58, Sept. 2006.

154. Z. Shen, Y. Xiao, L. Kang, W. Ma, L. Shi, L. Zhang, Z. Zhou, J. Yang, J. Zhong, D. Yang, L. Guo, G. Zhang, H. Li, Y. Xu, M. Chen, Z. Gao, J. Wang, L. Ren, and M. Li. Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. *Clin. Infect. Dis.*, Mar. 2020.

155. P. Skums, D. S. Campo, Z. Dimitrova, G. Vaughan, D. T. Lau, and Y. Khudyakov. Numerical detection, measuring and analysis of differential interferon resistance for individual hcv intrahost variants and its influence on the therapy response. *In silico biology*, 11(5):263–269, 2011.

156. P. Skums, Z. Dimitrova, D. S. Campo, G. Vaughan, L. Rossi, J. C. Forbi, J. Yokosawa, A. Zelikovsky, and Y. Khudyakov. Efficient error correction for next-generation sequencing of viral amplicons. *BMC Bioinformatics*, 13 Suppl 10:S6, June 2012.

157. P. Skums, N. Mancuso, A. Artyomenko, B. Tork, I. Mandoiu, Y. Khudyakov, and A. Zelikovsky. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows, 2013.

158. P. Skums, L. Bunimovich, and Y. Khudyakov. Antigenic cooperation among intrahost HCV variants organized into a complex network of cross-immunoreactivity. *Proc. Natl. Acad. Sci.*

U. S. A., 112(21):6653–6658, May 2015.

159. P. Skums, A. Artyomenko, O. Glebova, D. S. Campo, Z. Dimitrova, A. Zelikovsky, and Y. Khudyakov. Error correction of ngs reads from viral populations. *Computational Methods for Next Generation Sequencing Data Analysis*, 2016.

160. P. Skums, A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, L. Bunimovich, E. Costenbader, C. Sexton, S. O'Connor, G.-L. Xia, and Y. Khudyakov. QUENTIN: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics*, 34(1):163–170, Jan. 2018.

161. A. Sobel Leonard, M. T. McClain, G. J. D. Smith, D. E. Wentworth, R. A. Halpin, X. Lin, A. Ransier, T. B. Stockwell, S. R. Das, A. S. Gilbert, R. Lambkin-Williams, G. S. Ginsburg, C. W. Woods, and K. Koelle. Deep sequencing of influenza a virus from a human challenge study reveals a selective bottleneck and only limited intrahost genetic diversification. *J. Virol.*, 90(24):11247–11258, Dec. 2016.

162. D. A. Steinhauer and J. J. Holland. Rapid evolution of RNA viruses, 1987.

163. E. Thomson, C. L. C. Ip, A. Badhan, M. T. Christiansen, W. Adamson, M. A. Ansari, D. Bibby, J. Breuer, A. Brown, R. Bowden, J. Bryant, D. Bonsall, A. Da Silva Filipe, C. Hinds, E. Hudson, P. Klenerman, K. Lythgow, J. L. Mbisa, J. McLauchlan, R. Myers, P. Piazza, S. Roy, A. Trebes, V. B. Sreenu, J. Witteveldt, STOP-HCV Consortium, E. Barnes, and P. Simmonds. Comparison of Next-Generation sequencing technologies for comprehensive assessment of Full-Length hepatitis C viral genomes. *J. Clin. Microbiol.*, 54(10): 2470–2484, Oct. 2016.

164. A. Töpfer, O. Zagordi, S. Prabhakaran, V. Roth, E. Halperin, and N. Beerenwinkel. Probabilistic inference of viral quasispecies subject to recombination. *J. Comput. Biol.*, 20(2): 113–123, Feb. 2013.

165. A. Töpfer, T. Marschall, R. A. Bull, F. Luciani, A. Schönhuth, and N. Beerenwinkel. Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput. Biol.*, 10(3): e1003515, Mar. 2014.

166. V. Tsyvina, D. S. Campo, S. Sims, A. Zelikovsky, Y. Khudyakov, and P. Skums. Fast estimation of genetic relatedness between members of heterogeneous populations of closely related genomic variants. *BMC Bioinformatics*, 19(Suppl 11):360, Oct. 2018.

167. P. L. Tzou, P. Ariyaratne, V. Varghese, C. Lee, E. Rakhmanaliev, C. Villy, M. Yee, K. Tan, G. Michel, B. A. Pinsky, and R. W. Shafer. Comparison of an in vitro diagnostic Next-Generation sequencing assay with sanger sequencing for HIV-1 genotypic resistance testing. *J. Clin. Microbiol.*, 56(6), June 2018.

168. S. V. Vemula, J. Zhao, J. Liu, X. Wang, S. Biswas, and I. Hewlett. Current approaches for diagnosis of influenza virus infections in humans. *Viruses*, 8(4):96, Apr. 2016.

169. B. Verbist, L. Clement, J. Reumers, K. Thys, A. Vapirev, W. Talloen, Y. Wetzels, J. Meys, J. Aerssens, L. Bijnens, and O. Thas. ViVaMBC: estimating viral sequence variation in complex populations from illumina deep-sequencing data using model-based clustering. *BMC Bioinformatics*, 16:59, Feb. 2015.

170. B. M. P. Verbist, K. Thys, J. Reumers, Y. Wetzels, K. Van der Borght, W. Talloen, J. Aerssens, L. Clement, and O. Thas. VirVarSeq: a low-frequency virus variant detection pipeline for

illumina sequencing using adaptive base-calling accuracy filtering. *Bioinformatics*, 31(1): 94–101, Jan. 2015.

171. E. M. Volz, K. Koelle, and T. Bedford. Viral phylodynamics, 2013.

172. E. K. Wagner, M. J. Hewlett, D. C. Bloom, and D. Camerini. *Basic virology*, volume 3. Blackwell Science Malden, MA, 1999.

173. J. Wang, N. E. Moore, Y.-M. Deng, D. A. Eccles, and R. J. Hall. MinION nanopore sequencing of an influenza genome. *Front. Microbiol.*, 6:766, Aug. 2015.

174. W. Wang, X. Zhang, Y. Xu, G. M. Weinstock, A. M. Di Bisceglie, and X. Fan. High-resolution quantification of hepatitis c virus genome-wide mutation load and its correlation with the outcome of peginterferon-alpha2a and ribavirin combination therapy. *PloS one*, 9 (6):e100131, 2014.

175. R. L. Warren, G. G. Sutton, S. J. M. Jones, and R. A. Holt. Assembling millions of short DNA sequences using SSAKE, 2007.

176. T. M. Welzel, N. Bhardwaj, C. Hedskog, K. Chodavarapu, G. Camus, J. McNally, D. Brainard, M. D. Miller, H. Mo, E. Svarovskaia, I. Jacobson, S. Zeuzem, and K. Agarwal. Global epidemiology of HCV subtypes and resistance-associated substitutions evaluated by sequencing-based subtype analyses. *J. Hepatol.*, 67(2):224–236, Aug. 2017.

177. A. M. Wensing, V. Calvez, F. Ceccherini-Silberstein, C. Charpentier, H. F. Günthard, R. Paredes, R. W. Shafer, and D. D. Richman. 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.*, 27(3):111–121, 2019.

178. J. O. Wertheim, A. J. Leigh Brown, N. L. Hepler, S. R. Mehta, D. D. Richman, D. M. Smith,

and S. L. Kosakovsky Pond. The global transmission network of HIV-1. *J. Infect. Dis.*, 209 (2):304–313, Jan. 2014.

179. J. O. Wertheim, S. L. K. Pond, L. A. Forgione, S. R. Mehta, B. Murrell, S. Shah, D. M. Smith, K. Scheffler, and L. V. Torian. Social and genetic networks of hiv-1 transmission in new york city. *PLoS pathogens*, 13(1):e1006000, 2017.

180. K. Westbrooks, I. Astrovskaya, D. Campo, Y. Khudyakov, P. Berman, and A. Zelikovsky. HCV quasispecies assembly using network flows, 2008.

181. A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, and N. Nagarajan. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets, 2012.

182. C. K. Woods, C. J. Brumme, T. F. Liu, C. K. S. Chui, A. L. Chu, B. Wynhoven, T. A. Hall, C. Trevino, R. W. Shafer, and P. R. Harrigan. Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. *J. Clin. Microbiol.*, 50(6):1936–1942, June 2012.

183. C. Wymant, F. Blanquart, T. Golubchik, A. Gall, M. Bakker, D. Bezemer, N. J. Croucher, M. Hall, M. Hillebregt, S. H. Ong, O. Ratmann, J. Albert, N. Bannert, J. Fellay, K. Fransen, A. Gourlay, M. K. Grabowski, B. Gunsenheimer-Bartmeyer, H. F. Günthard, P. Kivelä, R. Kouyos, O. Laeyendecker, K. Liitsola, L. Meyer, K. Porter, M. Ristola, A. van Sighem, B. Berkhout, M. Cornelissen, P. Kellam, P. Reiss, C. Fraser, and BEEHIVE Collaboration. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with

shiver. *Virus Evol*, 4(1):vey007, Jan. 2018.

184. C. Wymant, M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, STOP-HCV Consortium, The Maela Pneumococcal Collaboration, and and The BEEHIVE Collaboration. PHYLOSCANNER: Inferring transmission from within- and Between-Host pathogen genetic diversity, 2018.

185. X. Yang, P. Charlebois, S. Gnerre, M. G. Coole, N. J. Lennon, J. Z. Levin, J. Qu, E. M. Ryan, M. C. Zody, and M. R. Henn. De novo assembly of highly diverse viral populations. *BMC Genomics*, 13:475, Sept. 2012.

186. X. Yang, P. Charlebois, A. Macalalad, M. R. Henn, and M. C. Zody. V-Phaser 2: variant inference for viral populations. *BMC Genomics*, 14:674, Oct. 2013.

187. R. J. F. Ypma, W. M. van Ballegooijen, and J. Wallinga. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–1062, Nov. 2013.

188. O. Zagordi, L. Geyrhofer, V. Roth, and N. Beerenwinkel. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J. Comput. Biol.*, 17(3):417–428, Mar. 2010.

189. O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12:119, Apr. 2011.

190. F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher. Population genomics of intrapatient hiv-1 evolution. *eLife*, Dec 2015. URL `https://elifesciences.org/articles/11282`.